Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

Bachelorarbeit

Harald Quitsch

Evaluation des Pentaho Open Source Business Intelligence Systems anhand von Big Data-Anwendungsszenarien

Fakultät Technik und Informatik
Department Informatik

Faculty of Engineering and Computer Science
Department of Computer Science

Harald Quitsch

Evaluation des Pentaho Open Source Business Intelligence Systems anhand von Big Data-Anwendungsszenarien

Bachelorarbeit eingereicht im Rahmen der Bachelorprüfung

im Studiengang Wirtschaftsinformatik
am Department Informatik
der Fakultät Technik und Informatik
der Hochschule für Angewandte Wissenschaften Hamburg

Betreuender Prüfer : Herr Prof. Dr. Olaf Zukunft
Zweitgutachter : Frau Prof. Dr. Ulrike Steffens

Abgegeben am 04.06.2015

Harald Quitsch

Thema der Bachelorarbeit
Evaluation des Pentaho Open Source Business Intelligence Systems anhand von Big Data-Anwendungsszenarien

Stichworte
Pentaho, Business Intelligence, Big Data, ETL, Open Source, OLAP, Reporting

Kurzzusammenfassung
Diese Ausarbeitung evaluiert das Open Source Business Intelligence System von Pentaho mittels eigens aufgestellter Big Data-Anwendungsszenarien. Neben einer Einführung der im Hintergrund stehenden Thematik, sowie näheren Erläuterung der Grundelemente dieses Systems wird auch auf die aktuelle Mitbewerbersituation eingegangen. Die anschließende Realisierung der Szenarien innerhalb der Anwendung wird abschließend ausführlich bewertet.

Harald Quitsch

Title of the paper
Evaluation of the Pentaho Open Source Business Intelligence System based on Big Data use cases

Keywords
Pentaho, Business Intelligence, Big Data, ETL, Open Source, OLAP, Reporting

Abstract
This paper evaluates the Open Source Business Intelligence System of Pentaho by using special Big Data use cases. Apart from introducing the background of this topic and also clarifying the basic elements of this system the current competitors are also discussed. Finally the subsequent implementation of these scenarios within the application is evaluated in detail.

Inhaltsverzeichnis

1. Einführung

1.1 Motivation und Zielsetzung

Spätestens seit Beginn des digitalen Zeitalters entwickelt sich auch die zugehörige Technik in rasanten Schritten. Dadurch ist in den letzten Jahren auch die Menge an verfügbaren Daten enorm angestiegen. Das sogenannte Internet der Dinge intensiviert diese Datenverfügbarkeit zusätzlich. Ziel ist es Alltagsgegenstände zunehmend mit dem Internet zu verbinden. Dadurch entwickeln sich diese zu interaktiven Gegenständen, welche in der Lage sind selbstständig Informationen zu sammeln und mit anderen Objekten oder mit dem Umfeld zu interagieren. Durchschnittlich besitzt in Deutschland bereits jetzt jede Person zwei Geräte, welche im Stande sind mit dem Internet zu kommunizieren (LfDI RLP 2015). Anzumerken ist hier jedoch, dass vermögendere oder auch technikaffine Personen durchaus über mehr Geräte verfügen und sich dadurch die Zahl derer, die womöglich keins besitzen relativiert.

Auch die Marktexperten prognostizieren in diesem Bereich einen gewaltigen Anstieg. Während Gartner[1] im Jahr 2020 von 33 Milliarden IP-fähigen Geräten weltweit ausgeht und die International Data Corporation (IDC)[2] mit 32 Milliarden Geräten rechnet, geht das Massachusetts Institute of Technology (MIT)[3] in Cambridge von 28 Milliarden Einheiten aus. Aktuell sind etwa 14 Milliarden Geräte vernetzt (Rohling 2014).

Es liegt auf der Hand, dass die Wirtschaft, die Gesellschaft, evtl. sogar jeder einzelne von diesen Daten profitieren kann. Die Big Player in diesem Geschäft – Google, Facebook, Amazon und Apple – haben sich in diesem Geschäftsbereich seit Jahren etabliert und ziehen trotz zum Teil negativer Kritik ihre Vorteile daraus. Doch auch andere Unternehmen werten Daten zunehmend aus und profitieren davon. Der Einsatz von Business Intelligence

[1] www.gartner.com
[2] www.idc.com
[3] www.mit.edu

Software vereinfacht dabei diese Analyse und Auswertung. Führungskräfte erhalten so relativ schnell und einfach elementare Informationen mit denen sie in der Lage sind für das Unternehmen in immer kürzeren Reaktionszeiten die richtigen Entscheidungen zu treffen. Eine mögliche Entwicklung in die falsche Richtung kann so schnell erkannt und korrigiert werden. Darüber hinaus können diese Auswertungen u.a. aber auch dabei helfen Geschäftsprozesse zu optimieren, Kosten zu senken, Risiken zu minimieren oder Kundenbeziehungen gewinnbringender zu gestalten.

Da die benötigte Software in diesem Anwendungsgebiet in der Regel mehrere Tausend Euro an Kosten verursacht, bleibt diese meist nur größeren Unternehmen vorbehalten. Kleine und mittelständische Unternehmen sind daher gezwungen sich nach kostengünstigeren Alternativen umzusehen um den Anschluss nicht zu verlieren und weiterhin Wettbewerbsfähigkeit zu bleiben. Der Open Source Markt z.B. bietet hier bereits einige Möglichkeiten.

Im Verlauf dieser Arbeit soll untersucht werden, ob die Open Source Lösung von Pentaho für solch Unternehmen aber auch für größere eine potenzielle Lösung darstellt. Um dieses Vorhaben umzusetzen werden praxisnahe Big Data-Anwendungsszenarien verwendet. Anhand gesetzter Ziele werden diese Szenarien bis ins Detail analysiert, ausgewertet und somit auch die Software auf den Prüfstand gestellt. Die erzielten Ergebnisse werden im Anschluss zusammengefasst und ausführlich bewertet.

1.2 Gliederung

Diese Ausarbeitung besteht aus insgesamt sechs Kapiteln. Neben der soeben dargestellten Einführung und der Zielstellung erfolgt im nächsten Kapitel eine Erläuterung der elementaren Grundlagen, welche für das Verständnis dieser Arbeit notwendig sind. Hierbei werden die Thematik Business Intelligence und die damit verbundenen Begriffe eingehend behandelt. Darüber hinaus werden die Grundelemente von Pentaho und dessen Mitbewerber beschrieben.
Das dritte Kapitel widmet sich den Big Data-Anwendungsszenarien, stellt diese dar, veranschaulicht deren zugrunde liegenden Daten und beschreibt dessen Hintergrund. Die Grundideen der Szenarien werden dann im vierten Kapitel innerhalb der Software realisiert. Anschließend erfolgt im fünften Abschnitt eine ausführliche Bewertung von Pentaho. Diese befasst sich u.a. mit der Realisierung der Szenarien innerhalb der Anwendung, dessen Funktionalität sowie dem Interaktionsaufwand des Nutzers. Das letzte Kapitel rundet die Arbeit mit einer Zusammenfassung ab und stellt einen möglichen Ausblick dar.

2. Grundlagen

Im Mittelpunkt dieses Kapitels stehen die Erläuterung der Begriffe Business Intelligence, Big Data und Open Source. Im weiteren Verlauf wird auf die Arbeitsweise des Systems von Pentaho eingegangen sowie auch ein kurzer Überblick über dessen Mitbewerber gegeben.

2.1 Business Intelligence Systeme

Der Begriff Business Intelligence (BI) beschreibt ein Konzept mit dessen Hilfe durch die Erfassung, Integration, Transformation, Speicherung, Analyse und Interpretation geschäftsrelevanter Daten neues Wissen generiert wird. Mit Hilfe dieses Wissens wird das Management in seinen Tätigkeiten unterstützt. Unter BI-Systeme versteht man in diesem Zusammenhang Softwarewerkzeugkästen in denen solche Datenbestände in großen Mengen integriert und ausgewertet werden (Hansen und Neumann 2009).

Typischerweise kommen derartige Systeme bei der Leistungsmessung interner Geschäftsprozesse, insbesondere für das Controlling, beim Aufbau eines Risikomanagements, beim Kennzahlenmanagement, bei der Unterstützung des Managements der Kundenbeziehungen (CRM) sowie bei der Analyse von Lieferantenbeziehungen zum Einsatz (Bange 2003).

2.1.1 Ablauf und Prozesse

Interne sowie externe Unternehmensdaten bilden hierbei die Grundlage für eine Auswertung. Ein wesentliches Problem stellt jedoch die Datenqualität dar. In den meisten Unternehmen sind die internen Daten historisch gewachsen und liegen so nicht einheitlich vor. Typische Qualitätsmängel sind fehlende, mehrfach vorkommende, falsch verknüpfte, falsch definierte sowie auch inhaltlich falsche Daten. Um diese Mängel zu beseitigen durchlaufen die Daten einen Extract-Transform-Load-Prozess (ETL-Prozess). Anschließend werden sie in ein Data-Warehouse eingepflegt und dort zur Analyse über kurze, mittlere und längere Zeiträume gespeichert (siehe Abbildung 2.1).

Die Analyse kann dabei nach unterschiedlichen Kriterien erfolgen. Für gewöhnlich werden hier Dimensionen wie Zeit, Region, Produkt, Lieferant, oder Kunde eingerichtet, aber auch andere sind denkbar (Hansen und Neumann 2009).

Die wesentlichen Kernapplikationen, die eine Entscheidungsfindung zielgerichtet unterstützen sind das Online Analytical Processing (OLAP), das Executive Information System (EIS) und das Data Mining (Kemper, Baars und Mehanna 2010).

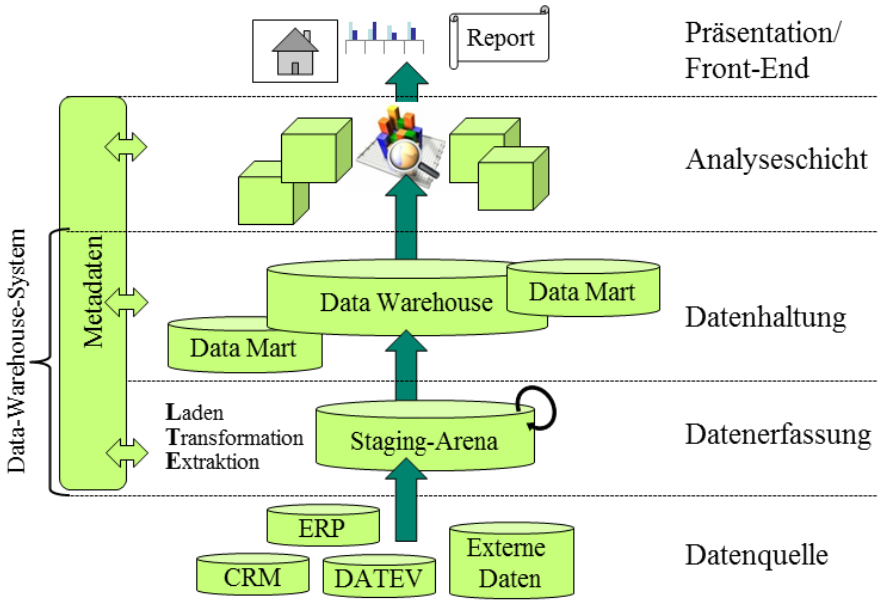


Abbildung 2.1 – Ordnungsrahmen (Rödl Dynamics AG 2014)

2.1.1.1 ETL-Prozess

Da es in der Regel sehr aufwendig ist vorhandene Daten in einen konsistenten und homogenen Zustand zu überführen, kommt diesem Prozessschritt eine besondere Bedeutung. Dieser wird daher in den drei Phasen Extract, Transform und Load gegliedert.
In der ersten Phase werden selektierte Daten aus verschiedenen Quellen, wie z.B. ERP-Systeme, E-Mails sowie Datenbanken, extrahiert und zwischengespeichert. Anschließend unterlaufen diese einer syntaktischen sowie semantischen Transformation. Die syntaktische Transformation bezieht sich dabei auf rein formale Aspekte. So werden hier alle Daten dahingehend modifiziert, dass sie die für das Zielsystem notwendige Syntax aufweisen. In der semantischen Transformation werden die Daten hinsichtlich inhaltlicher Aspekte überprüft und ggf. angepasst. So werden sie beispielsweise aggregiert, gefiltert oder gar mit Zusatzinformationen angereichert.
Nach Abschluss der zweiten Phase werden die Daten in ein Data-Warehouse geladen. Um die Daten auf dem aktuellen Stand zu halten, wird der ETL-Prozess je nach Anwendungsfall periodisch durchgeführt. Dies kann z.B. täglich, monatlich oder direkt nach dem Erzeugen neuer Daten geschehen (Bange 2003).

2.1.1.2 Online Analytical Processing

OLAP steht für einen dynamischen, flexiblen und interaktiven Zugriff auf große Datenmengen (Bauer und Günzel 2013).

Da die Daten aus mehreren Geschäftsfeldern stammen und so verschiedene Aspekte wiederspiegeln spricht man auch von multidimensionalen Daten. Diese werden charakteristisch in so genannte Cubes (Würfel) dargestellt (siehe Abbildung 2.2).

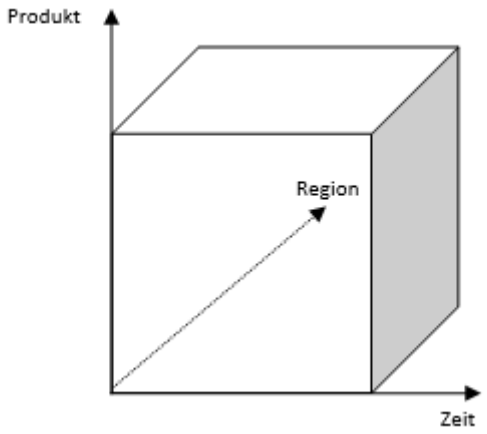


Abbildung 2.2 – Cube angelehnt an (Hansen und Neumann 2009)

Grundsätzlich existieren drei verschiedene Arten des OLAP welche sich im jeweiligen Speicherkonzept unterscheiden. Werden die multidimensionalen Daten vollständig im Cube gespeichert, so spricht man vom Multidimensional Online Analytical Processing (MOLAP). Erfolgt die Speicherung jedoch in einer zugrunde liegenden relationalen Datenbank, so werden im Cube lediglich die multidimensionalen Definitionen festgelegt. Diese Variante wird als Relational Online Analytical Processing (ROLAP) bezeichnet. Die dritte Art der Speicherung stellt eine Mischform von MOLAP und ROLAP dar und wird als Hybride Online Analytical Processing (HOLAP) definiert (Azevedo, et al. 2009).

Mit Hilfe verschiedener Funktionen erfolgt dann der Zugriff auf die gewünschten Daten. Die meist genutzten Operationen sind die Pivotierung, der Drill Down bzw. Roll Up, Slice & Dice sowie Split & Merge (Kemper, Baars und Mehanna 2010).

In vielen Fällen sind bereits zwei Dimensionen für betriebliche Analysen ausreichend. Solch eine Sicht stellt z.B. eine Seite des Cubes dar. Mit Hilfe der Pivotierung, oftmals auch als Rotation bezeichnet, wird der Cube um eine Achse gedreht. Durch diesen Vorgang wird eine andere Kombination von zwei Dimensionen sichtbar (siehe Abbildung 2.3).

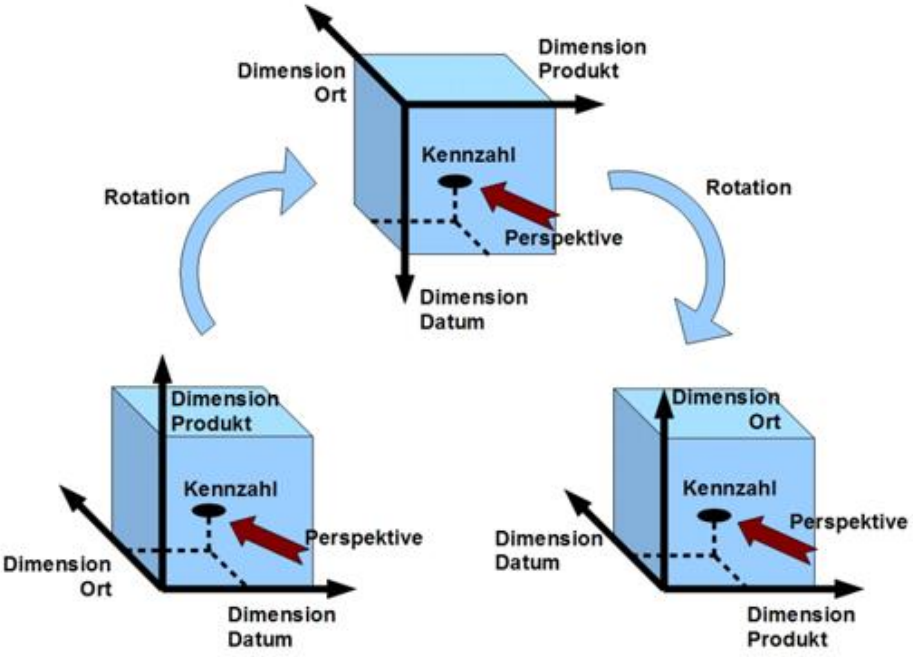


Abbildung 2.3 – Pivotierung (DATACOM Buchverlag GmbH 2015)

Befinden sich die dargestellten Informationen in einem aggregierten Zustand, können diese durch einen Drill Down wieder aufgefächert werden um mehr Details zu erhalten. Umgekehrt lassen sich die Werte auch wieder verdichten. Diese Verringerung des Detaillierungsgrad wird als Roll Up bezeichnet (siehe Abbildung 2.4).

	JAHRESUMSATZ
PRODUKT A	560.000 €
PRODUKT B	400.000 €
PRODUKT C	800.000 €
PRODUKT D	480.000 €

Drill Down

Roll Up

	JAHRESUMSATZ
PRODUKTGRUPPE ABCD	2.240.000 €

Abbildung 2.4 – Drill Down & Roll Up angelehnt an (Kemper, Baars und Mehanna 2010)

Mit Hilfe von Slice & Dice können Anwender einen bestimmten Ausschnitt der im Cube aggregierten Daten bedarfsgerecht filtern. Ein Slice stellt dabei eine Scheibe dar, die aus dem Cube entnommen wird. Der Anwender kann so z.B. alle Daten in Bezug zu einem Produkt einsehen (siehe Abbildung 2.5). Ein Dice dagegen ist ein mehrdimensionaler Ausschnitt (Teilwürfel) des Cubes. Dieser entsteht, indem in unterschiedlichen Dimensionen die jeweiligen Dimensionselemente in ihrer Menge eingeschränkt werden. Dieser neue Teilwürfel kann im Anschluss weiter extrahiert oder verarbeitet werden.

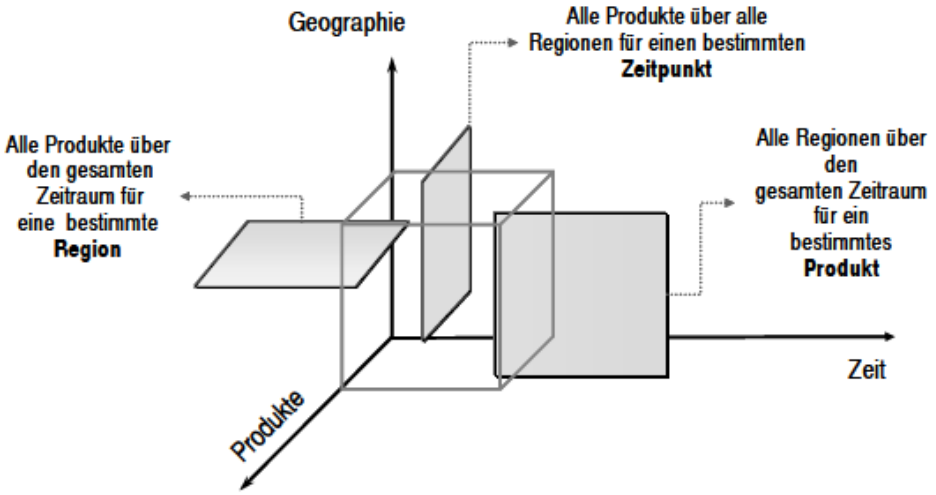


Abbildung 2.5 – Slice Funktion (Kemper, Baars und Mehanna 2010)

Mittels der Split Funktion kann ein Wert nach Elementen einer weiteren Dimension weiter aufgegliedert werden. Dies ermöglicht einen höheren Detaillierungsgrad der dargestellten Informationen (siehe Abbildung 2.6). Die Merge Funktion entfernt den Einschub der zusätzlichen Dimension wieder, wodurch die Granularität der Darstellung reduziert wird.

Geographie ↓ Produkte ↓ → Zeit

		1.Quartal	2. Quartal	3. Quartal
Stadtmitte	Produkt A	12.000	12.000	12.000
	Produkt B	8.000	10.000	8.000
	Produkt C	16.000	16.000	16.000
West	Produkt A	10.000	8.000	10.000

Abbildung 2.6 – Split Funktion (Kemper, Baars und Mehanna 2010)

2.1.1.3 Executive Information System

Ein Executive Information System (EIS) bezeichnet ein dialog- und datenorientiertes Softwaresystem, welches speziell für Fach- und Führungskräfte zugeschnitten ist. Durch die benutzerfreundliche Darstellung von aktuellen und entscheidungsrelevanten internen sowie externen Daten werden die Manager gezielt bei ihren Aufgaben unterstützt.

Dabei sind EIS immer unternehmungsspezifisch aufgebaut und werden neben des Kontrollprozesses hauptsächlich in den ersten Phasen des Planungs- und Entscheidungsprozesses eingesetzt. Weiterführende integrierte Funktionen, wie z.B. die Kommunikationsfähigkeit mittels Mailing-Systemen, die ebenfalls vorhandene Drill-Down- oder die Prognose-Methode steigern die Nutzbarkeit dieses Systems (Gabriel 2013).

2.1.1.4 Data-Mining

Mit dem Begriff Data-Mining verbindet man einen softwaregestützten, systematischen Prozess aus großen Datenbeständen neue Zusammenhänge, Muster und Trends zu erkennen. Dafür werden Methoden der Statistik und des maschinellen Lernens herangezogen, welche meist den drei Kategorien Klassifikation, Regression und Segmentierung zugeordnet werden.
Die Klassifikation beschäftigt sich hierbei mit qualitativen Merkmalen. Auf Basis vorhandener Variablen ist es das Ziel, ein neues Merkmal zu kreieren. So verfügt man z.B. bei Bankkunden über Informationen des Einkommens, des Vermögens sowie der Kredithistorie und kann die Kunden daraufhin in das neue Merkmal Bonitätsstufe einteilen.
Dagegen prognostiziert die Regression ein quantitatives Merkmal. So wird hier vergleichsweise die Nachfrage von Lebensmitteln aufgrund der Abhängigkeit früherer Käufe, dem Wochentag und der Werbeschaltungen bestimmt.
In der dritten Kategorie, auch als Clustering bezeichnet, werden Ausprägungsgruppen aufgrund von Ähnlichkeiten erstellt. So segmentiert man beispielsweise den Markt von Kunden aufgrund von Präferenzen (Hansen und Neumann 2009).

2.2 Big Data

Obwohl für den Begriff Big Data noch keine verbindliche Definition existiert, wird dieser oftmals mit den drei V's – Volume, Variety und Velocity – charakterisiert. Diese wurden bereits im Jahr 2001 von Douglas Laney, dem damaligen Analysten der Meta Group – heute Gartner[4] – eingeführt (Laney 2012).

[4] www.gartner.com

Volume beschreibt die in diesem Zusammenhang umfangreiche und kontinuierlich anfallende Datenmenge. Die Entstehung dieser hat mittlerweile eine Eigendynamik angenommen, die weder gesteuert noch kontrolliert werden kann. Jede Person, sowohl beruflich als auch privat, ist in gleicher Weise Nutzer und Erzeuger von Daten. Sich diesem zu entziehen ist nahezu unmöglich. Das Internet und Fernsehen, Haushaltsgeräte jeglicher Art sowie mobile Geräte mit Geo-Daten verschmelzen mit dem Fortschritt der Technologie immer weiter miteinander und kreieren so unaufhörlich neue Daten (Bachmann, Kemper und Gerzer 2014).

Variety kennzeichnet die in diesem Datenpool enthaltene Heterogenität. Die zusammengeflossenen Daten haben ihren Ursprung in den unterschiedlichsten Quellen, wodurch sich auch deren Strukturen voneinander unterscheiden. Differenziert wird hier zwischen strukturierten (z.B. relationale Datenbanken), halbstrukturierten (z.B. E-Mails) und unstrukturierten Daten (z.B. Texte oder Multimedia). Die unstrukturierten Daten, welche den größten Teil dieser Datenmenge ausmachen, lassen sich wiederum in drei weitere Unterkategorien zerlegen (Klein, Tran-Gia und Hartmann 2013):

- Daten aus der Kommunikation zwischen Personen, z.B.:
    - Social Networks

- Daten aus der Kommunikation zwischen Personen und Diensten oder Maschinen, z.B.:
    - E-Commerce Anwendungen
    - Nutzung von Geldautomaten

- Daten aus der Kommunikation zwischen Maschinen, z.B.:
    - Sensordaten
    - GPS-Informationen
    - Überwachungsbilder

Velocity beschreibt die Geschwindigkeit in der diese Datenmenge entsteht, aber auch den Zeitrahmen von der Datenankunft bis hin zur Verarbeitung im Unternehmen. Wie man der Abbildung 2.7 entnehmen kann, prognostizieren die Marktbeobachter der IDC und der EMC Corporation (EMC)[5] alle zwei Jahre eine Verdopplung dieser Datenmenge, so dass diese im Jahr 2020 sogar bei ca. 40 Zettabyte (ZB) liegen soll (Die Welt 2015).

[5] www.emc.com

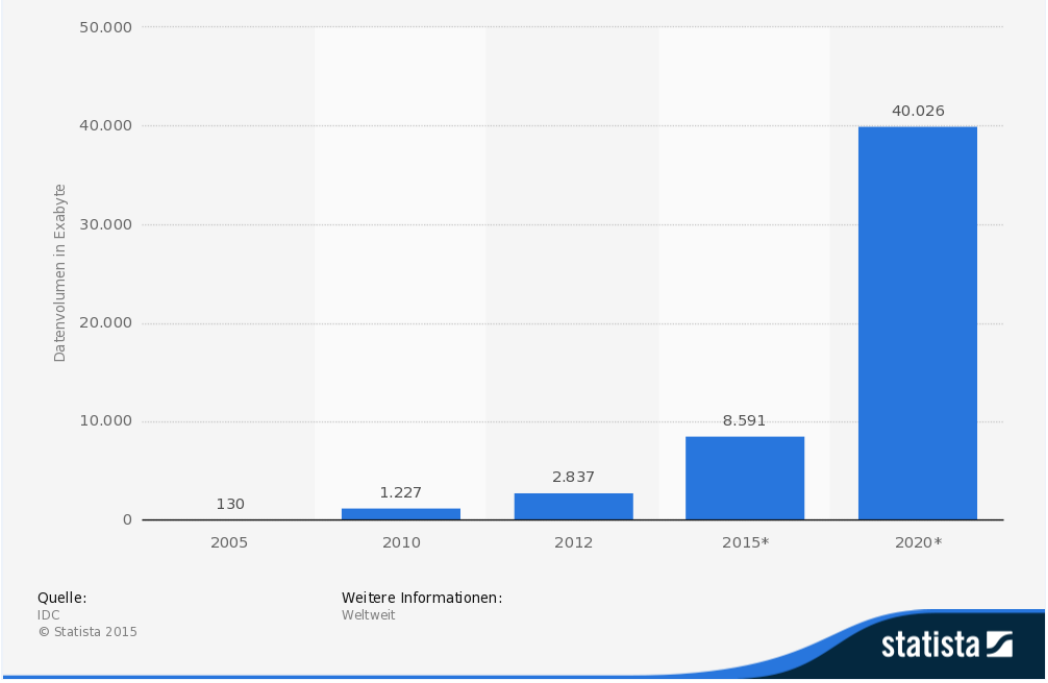


Abbildung 2.7 – Prognose der weltweit generierten Datenmenge (Statista GmbH 2015)

Angesichts dieser Entwicklung spielt auch die Datenqualität eine zunehmend wichtige Rolle. Diese weitere Eigenschaft wird häufig als das vierte V – Veracity – zur Definition von Big Data hinzugenommen (Jagadish, et al. 2014). Hierbei geht es aber auch um die Vertrauenswürdigkeit der Daten bzw. der jeweiligen Quellen. Aus diesem Grund sollte die für die Weiterverarbeitung verwendete Datenmenge vorab untersucht und ggf. bereinigt werden. Im Hinblick auf die Datenqualität haben die Autoren Wang und Strong 1500 datenverarbeitende Unternehmen befragt und aus 179 Qualitätsmerkmalen die 15 wichtigsten zusammengestellt (siehe Tabelle 2.1), die für diesen Schritt herangezogen werden können.

MERKMALSKLASSE	QUALITÄTSMERKMAL
INTRINSISCHE DATENQUALITÄT	Glaubhaftigkeit
	Genauigkeit
	Objektivität
	Reputation
KONTEXTUELLE DATENQUALITÄT	Mehrwert
	Relevanz
	Aktualität
	Datenmenge
	Vollständigkeit
REPRÄSENTATIONELLE DATENQUALITÄT	Interpretierbarkeit
	Verständlichkeit
	Konsistenz
	Präzision
ZUGRIFFSQUALITÄT	Verfügbarkeit
	Zugriffssicherheit

Tabelle 2.1 – Datenqualitätsmerkmale angelehnt an (Wang und Strong 1996)

2.2.1 Big Data Life Cycle

Damit aus dieser Datenmenge ein Nutzen erschlossen werden kann, durchläuft diese einen mehrstufigen, iterativen Prozess, der von den Autoren des Artikels „Big Data and Its Technical Challenges“ (Jagadish, et al. 2014) als Big Data Life Cycle bezeichnet wird.

Dieser beginnt mit der Erfassung und Filterung der relevanten Daten, da eine Speicherung aller aufgrund der Größe oft nicht möglich ist. Da die nun vorliegende Datenmenge verschiedenste Datenformate enthält und so nicht einheitlich analysiert werden kann, werden die wichtigen Informationen aus diesen Formaten extrahiert und in ein einheitliches Format gebracht. Ebenso werden in diesem Schritt womöglich fehlerhafte oder unzuverlässige Daten entfernt. Diese nun bereinigten Daten werden in der nächsten Phase im Unternehmen gespeichert und anschließend analysiert. Die Ergebnisse dieser Analyse werden letztendlich von einem Entscheidungsträger interpretiert und auch hinterfragt, um die in den Analysen getroffenen Annahmen auf ihre Richtigkeit hin zu prüfen und mögliche Softwarefehler auszuschließen. Dieser Big Data Life Cycle (siehe Abbildung 2.8) kann als ein nicht endend wollender Prozess angesehen werden, da im Anschluss immer wieder neue und aktuellere Daten erfasst werden die ebenfalls diesen Zyklus durchlaufen.

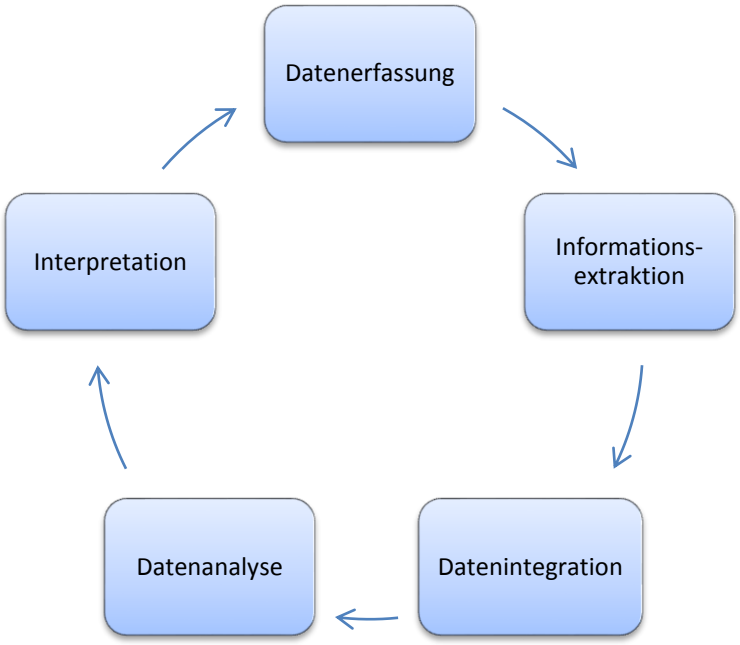


Abbildung 2.8 – Big Data Life Cycle anhand (Jagadish, et al. 2014)

2.3 Open Source

Unter dem Begriff Open Source (OS) versteht man den freien Zugang zum Quellcode von Softwareprogrammen. Die Nutzung dieser Programme ist kostenlos und jede Person darf diese aus dem Internet laden und verbreiten. Damit eine Software aber auch offiziell als Open Source anerkannt wird, müssen folgende Kriterien erfüllt sein (Open Source Initiative 2014a):

1. **Freie Weiterverbreitung:** Die Lizenz darf niemanden in der Weitergabe einschränken. Darüber hinaus dürfen auch keinerlei Lizenzgebühren oder andere Beträge erhoben werden.
2. **Quellcode:** Das Programm muss den Quellcode in einer verständlichen Programmiersprache enthalten und eine Verteilung erlauben. Ist dies nicht der Fall, muss der Quellcode öffentlich, ohne weitere Kosten und vorzugsweise über das Internet zugänglich sein.
3. **Weiterführende Arbeiten:** Die Lizenz muss Modifikationen gestatten und es erlauben diese modifizierte Version unter derselben Lizenz wie die der Original-Software zu veröffentlichen.

4. **Unversehrtheit des Originalcodes:** Wird ein veränderter Quellcode verbreitet, so müssen die entsprechenden Änderungen so gekennzeichnet werden, dass deutlich wird welcher Code-Teil aus dem Originalcode stammt.
5. **Keine Diskriminierung von Personen oder Gruppen:** Die Lizenz darf keine einzelne Person oder Personengruppen diskriminieren.
6. **Keine Einschränkung der Anwendungsbereiche:** Die Lizenz darf kein bestimmtes Einsatzgebiet einschränken.
7. **Verbreitung der Lizenz:** Die Rechte einer Software gehen auf alle Personen über, die diese verwenden. Eine zusätzliche Lizenz muss für diesen Zweck nicht erworben werden.
8. **Die Lizenz darf nicht für ein bestimmtes Produkt gelten:** Werden aus Softwarepaketen einzelne Programme weiter verbreitet, so gelten für diese dieselbe Lizenz wie für das Softwarepaket.
9. **Die Lizenz darf keine andere Software beeinträchtigen:** Die Lizenz darf keine anderen Programme, die z.B. in einem Softwarepaket enthalten sein können, einschränken.
10. **Die Lizenz muss technikneutral sein:** Die Lizenz darf nicht nur für eine bestimmte Technologie oder Schnittstelle gelten.

2.3.1 Lizenzmodelle

Mittlerweile gibt es an die 80 Lizenzen denen Open Source Software (OSS) unterliegt und die sich auch in verschiedene Kategorien einteilen lassen (Open Source Initiative 2014b). Die wichtigsten und auch bekanntesten Modelle sind jedoch die General Public Licence (GPL) und die Berkeley Software Distribution (BSD) - Licence (Hansen und Neumann 2009).

Mit der GPL wird sichergestellt, dass eine Software bei Weiterentwicklung bzw. Veränderung ebenfalls unter dieser Lizenz gestellt werden muss. Dieses Prinzip wird auch als strenges Copyleft bezeichnet. So wird gewährleistet, dass die Software stets kostenlos bleibt.

Die BSD-Lizenz dagegen besitzt kein Copyleft und besagt, dass die Software frei verwendet, verändert, kopiert und verbreitet werden kann, solange der Copyright-Hinweis des ursprünglichen Autors mitgeliefert wird. Demnach ermöglicht diese Lizenz die Entwicklung kommerzieller Software, welche auf OSS basiert und unter einer anderen, beliebigen Lizenzbedingung in den Umlauf gebracht werden kann. Eine Offenlegung des Quellcodes ist dann auch nicht mehr verpflichtend.

Neben diesen beiden gibt es auch solche, die über ein eingeschränktes Copyleft verfügen. Diese Einschränkung vereinfacht die Kombination von Software mit unterschiedlichen Lizenzen. Ein bekanntes Modell ist hier die Lesser General Public Licence (LGPL), welche

speziell für Programmbibliotheken gedacht ist und die Verlinkung dieser Bibliotheken sowohl mit herkömmlicher als auch mit OSS gestattet (ifrOSS 2015).

2.4 Pentaho

Die Pentaho BI-Suite ist ein Produkt der Pentaho Corporation[6] und entstand 2004 durch eine Zusammenfassung der verschiedenen Projekte JFreeReport, Kettle, Mondrian sowie Weka, welche allesamt aus dem im Kapitel 2.3 definierten Open Source Bereich stammen (it-novum GmbH 2009). Insgesamt unterliegt die Suite der im Abschnitt 2.3.1 beschriebenen General Public Licence[7].

Laut Pentaho hatte man den Wunsch den Geschäftsanalysemarkt positiv zu verändern, der bis dato von den kommerziellen Produkten der Großanbieter beherrscht wurde. Mittlerweile hat sich Pentaho eine starke Marktposition erarbeitet und führt Geschäftsanalysen für Tausende von Unternehmen und Software-Herstellern auf der ganzen Welt aus. Der Hauptfirmensitz liegt in Florida, weitere Niederlassungen sind in Kalifornien sowie in Europa zu finden (Pentaho Corporation 2014).

Zur Verfügung stehen zum einen die Pentaho Community Edition (CE) und zum anderen die Pentaho Enterprise Edition (EE). Die CE, auf welche in diesem Kapitel auch noch näher eingegangen wird, ist für jedermann per Download kostenfrei zugänglich. Die EE dagegen muss kostenpflichtig erworben werden, wobei diese vorab 30 Tage lang kostenfrei getestet werden kann. Im Gegenzug werden ein professioneller Support, insbesondere bei der Installation, sowie weitere Programm-Features zur Verfügung gestellt.

Die Suite verfügt über einen modularen Aufbau der einzelnen Module, welche im Rahmen dieser, aber auch unabhängig voneinander verwendet werden können. Insgesamt werden so Möglichkeiten für den Datenzugriff bzw. die Datenintegration, die Datenerkennung, die Datenanalyse und die Datenvisualisierung angeboten.

Die nachstehende Abbildung 2.9 verdeutlicht den Aufbau der Suite:

[6] www.pentaho.com
[7] http://community.pentaho.com/faq/platform_licensing.php

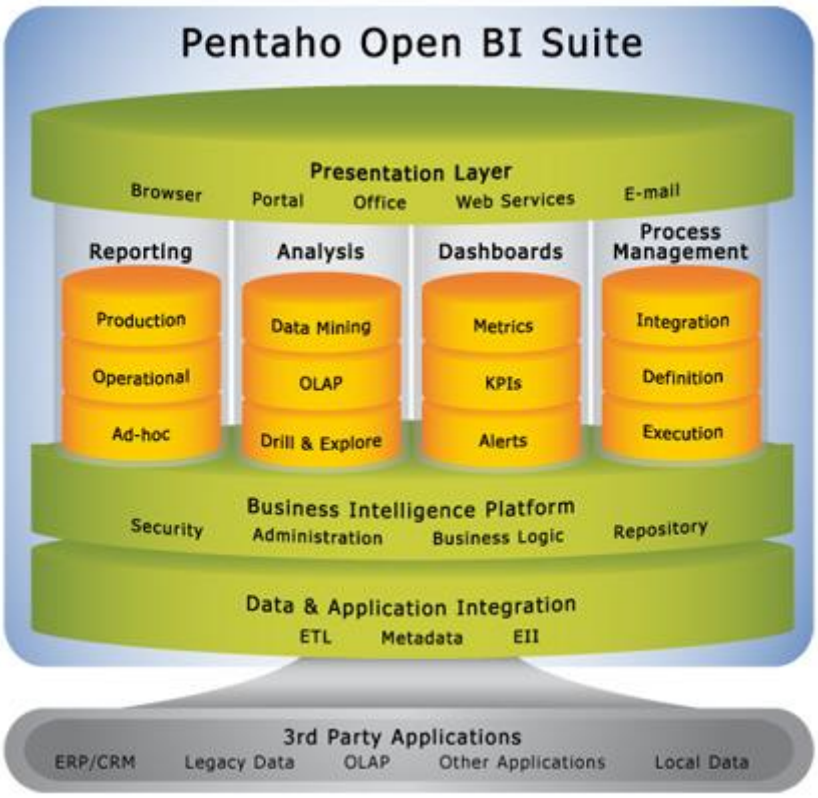


Abbildung 2.9 – Funktionelle Architektur der Pentaho Suite (Pentaho Corporation 2014)

2.4.1 Pentaho Data Integration

Mit Pentaho Data Integration werden die Daten aus allen möglichen Quellen vorbereitet. Dahinter verbirgt sich ein ETL-Tool, welches aus dem Open Source-Projekt Kettle hervorgegangen ist und aus vier Einzelapplikationen besteht (Pentaho DI Documentation 2014).

- **Spoon** stellt die grafische Benutzerschnittstelle dar, mit der Datentransformationen und Jobs erstellt werden können. Jobs beinhalten unter anderem mehrere Datentransformationen.

- **Pan** ist für die Datentransformation, welche den Import und Export aus verschiedenen Quellen umfasst, zuständig und führt eine Vielzahl von Funktionen wie Lesen, Manipulieren und Schreiben aus. Die durchzuführenden Transformationen kommen hier in dem Extensible Markup Language (XML) -Format an und werden in bestimmten Intervallen als Stapel verarbeitet.

- **Kitchen** ist für die Stapelverarbeitung von Job-Ketten zuständig, die ebenfalls in bestimmten Intervallen erfolgt.

- **Carte** ist ein einfacher Webserver mit dessen Hilfe Datentransformationen sowie Jobs per remote ausgelöst, beobachtet und gestoppt werden können.

Mit Hilfe von Plug-Ins ist Pentaho Data Integration jederzeit erweiterbar. Für diesen Zweck existieren neben freien auch kostenpflichtige Plug-Ins.

2.4.2 Pentaho Analysis

Die Analyse großer Datenmengen erfolgt zum einen mit dem OLAP Mondrian, der aus dem selbigen Open Source-Projekt entstanden ist und zum anderen mit Hilfe des Pentaho Data Mining.

Mit Hilfe von Mondrian sind, unabhängig der Datenmenge, interaktive Anfragen in Echtzeit möglich. Wesentliche Bestandteile von Mondrian sind die Schema-Workbench, die Mondrian-Webanwendung und der Aggregation-Designer. Mittels Drag & Drop werden in der Schema-Workbench Cubes definiert und als XML-Datei gespeichert. Die Mondrian-Webanwendung ermöglicht die Durchführung von OLAP-Analysen (vgl. Abschnitt 2.1.1.2) und stellt die Ergebnisse grafisch in Form von Tabellen oder Diagrammen dar. Mit dem integrierten OLAP-Navigator lassen sich die Dimensionen des Cubes frei kombinieren und auch die Auswahl der gewünschten Daten frei bestimmen. Darüber hinaus können die Cubes auch in Excel mit Hilfe von Pivot-Tabellen ausgewertet werden. Der Aggregation Designer steigert die Leistung von OLAP-Analysen und erstellt Tabellen, die der Datenbank den Aggregationsprozess ersparen (Pentaho Mondrian 2014).

Für die vorausschauende Analyse gibt es das Pentaho Data Mining (vgl. Abschnitt 2.1.1.4), welches auf dem Open Source-Projekt WEKA basiert. Angeboten wird hier eine breite Palette von Klassifikationen, Regressionen, Assoziationsregeln und Clustering-Algorithmen. Das Tool bietet als Oberfläche den Explorer sowie den Experimenter an. Im Explorer gibt der Anwender die Datenquelle an, aus der die Daten importiert werden sollen. Darüber hinaus werden hier auch Daten geändert und gefiltert um anschließend Algorithmen ausführen zu können.
Im Experimenter werden Lernansätze in Form von Experimenten angelegt und getestet. Hierfür gibt der Anwender die Datenquelle an und wählt die gewünschten Algorithmen aus. Das anschließende Ergebnis wird dann in einem eigenen Fenster dargestellt (Pentaho DM Documentation 2014).

2.4.3 Pentaho Reporting

Für die Berichterstellung bietet Pentaho den Report Designer an. Es ist eine Weiterführung des Open Source-Projektes JFreeReport. Dem Anwender steht hier eine Oberfläche zur Verfügung mit welcher Berichte in verschiedenen Varianten generiert werden können. Die gewünschten Komponenten werden hierfür einfach per Drag & Drop an die gewünschte Stelle platziert.
Die Ausgabe der Berichte kann in den Formaten Hypertext Markup Language (HTML), Excel, Comma Separated Values (CSV), Portable Document Format (PDF) sowie Rich Text Format (RTF) erfolgen und auf der BI-Plattform veröffentlicht werden (Pentaho Reporting 2014).

2.4.4 Pentaho Dashboard

Ein Dashboard stellt dem Anwender wichtige Informationen, wie z.B. ermittelte Kennzahlen, in grafischer Form zur Verfügung. In der CE steht jedoch keine Oberfläche für solch eine Erstellung bereit. Dennoch ist es möglich diese mit den integrierten Bibliotheken des Projekts Community Dashboards Framework manuell zu realisieren (Webdetails 2014).

2.5 Mitbewerber

Da BI für die Unternehmensführung zunehmend an Bedeutung gewonnen hat, existiert auch eine Vielzahl von Anwendungen in diesem Bereich. Den Großteil des Marktes teilen sich jedoch Firmen wie Microsoft, IBM, SAP und Oracle mit ihren kommerziellen Lösungen. Dennoch stellen die Open Source Lösungen der Mitbewerber mittlerweile eine ernsthafte Konkurrenz dar und werden auch schon von namhaften Unternehmen eingesetzt. Neben der Vielzahl an Anbietern, die Lösungen für einzelne Bereiche eines BI-Systems anbieten, gibt es auch solche die wie Pentaho über ein Gesamtpaket verfügen. Hervorzuheben sind hier der deutsche Anbieter Jedox[8] sowie der amerikanische Anbieter Jaspersoft[9].
Der Artikel „Business Intelligence mit Open Source“ von Ilkem Güclü und Stefan Müller (Güclü und Müller 2010) vergleicht unter anderem diese beiden Lösungen miteinander.

So verfügen beide Anbieter über eine kostenlose Community-Version, welche sich jedoch kostenpflichtig upgraden lässt. Im Gegenzug werden weitere Features zur Verfügung gestellt sowie ein Support angeboten.
Jedox orientiert sich dabei eher an Fachanwendern, indem das Analyseinstrument über ein Excel-Frontend angeboten wird. Jaspersoft dagegen verfolgt einen plattformorientierten Ansatz mittels modularer Werkzeuge.

[8] www.jedox.com
[9] www.jaspersoft.com

Für die Datenintegration verwendet Jedox den Palo ETL-Server, der aus einer Java-Anwendung besteht und als Webdienst zur Verfügung gestellt wird. Mit Hilfe von ETL-Prozessen wird die sogenannte Palo-Datenbank befüllt. Die im nächsten Schritt notwendige Aufbereitung der Daten erfolgt mittels MOLAP und einer In-memory-Technologie. So werden die Daten in einem bestimmten Datenbankformat und auch im Arbeitsspeicher für die Analyse bereitgehalten. Dies ermöglicht eine hohe Abfragegeschwindigkeit sowie ein direktes zurückschreiben der Daten über das Frontend. Das Ergebnis der analysierten Daten wird in Berichten präsentiert, welche jedoch nur mittels der Excel-Funktionen formatiert und grafisch dargestellt werden können. Auch eine Verteilung dieser ist nur eingeschränkt über einen File-Server oder per E-Mail möglich. Durch Integration des Worksheet-Servers lassen sich diese Einschränkung jedoch aufheben und man kann die Berichte zusätzlich über eine Weboberfläche verteilen sowie auch spezielle Grafiken nutzen.

Im Gegensatz dazu nutzt Jaspersoft für die Datenintegration die Anwendung JasperETL mit dessen Hilfe ETL-Prozesse per Drag & Drop zusammengestellt werden können. Auch die Datenanalyse unterscheidet sich. Statt eines MOLAP kommt ein ROLAP zum Einsatz. So werden die Daten in relationalen Datenbanken gespeichert und durch besondere Schemata aufbereitet. Für die Ergebnisdarstellung kann ein Ad-hoc-Reporting im Browser durchgeführt werden, wodurch Daten nach Belieben gefiltert, sortiert und in unterschiedlichen Detaillierungsgraden betrachtet werden können. Mit dem Berichtsdesigner können darüber hinaus Vorlagen erstellt werden, die regelmäßig mit Daten befüllt werden. Fertige Berichte können ebenfalls über eine Weboberfläche oder einen Fileserver verteilt werden.

Die nachfolgende Tabelle 2.2 gibt einen Überblick der einzelnen Elemente der jeweiligen Anbieter.

	Pentaho	**Jedox**	**Jaspersoft**
Datenintegration	Pentaho Data Integration (ETL-Tool)	Palo ETL-Server	JasperETL (Drag & Drop)
Datenanalyse	Pentaho Analysis (OLAP + Pentaho Data Mining)	MOLAP + In-memory-Technologie	ROLAP
Ergebnisdarstellung	Pentaho Reporting (Berichte) & Pentaho Dashboard	Excel-Frontend + Weboberfläche (Berichte)	Plattformorientiert mittels modularer Werkzeuge (Ad-hoc-Reporting, Berichtsdesigner)

Tabelle 2.2 – Ausschnitt der OS BI-Suite Anbieter

3. Anwendungsszenarien

In diesem Kapitel wird im Detail auf die Big Data - Anwendungsszenarien eingegangen, welche im späteren Verlauf in das Business Intelligence System (vgl. Abschnitt 2.1) von Pentaho integriert werden.

3.1 Datenbeschreibung

Gemäß dem im Abschnitt 2.2 beschriebenen Begriff Big Data liegt hier eine von der New York City Taxi & Limousine Commission (TLC) bereitgestellte Datenmenge aus dem Jahr 2013 zu Grunde.

Die in 1971 gegründete TLC lizensiert und reguliert über 50.000 Fahrzeuge und rund 100.000 Fahrer und ist damit die aktivste Taxi- und Limousinenlizenzaufsichtsbehörde in den Vereinigten Staaten von Amerika. Darüber hinaus führt sie dreimal im Jahr Sicherheits- und Emissionskontrollen der 13.637 Medaillon Taxis durch und inspiziert alle zwei Jahre die durch die TLC lizensierten Mietfahrzeuge. Das Ziel der TLC ist es allen New Yorkern sowie auch Besuchern einen sicheren und effizienten Zugang zum Beförderungsnetzwerk zu ermöglichen, so dass ein gutes Reiseerlebnis gewährleistet werden kann (TLC 2014).

Insgesamt beträgt die Datenmenge 45,20 Gigabyte (GB) und unterteilt sich in Tarif[10]- sowie Reisedaten[11]. Beide Datengruppen enthalten jeweils 12, nach Monaten abgegrenzte, Datenblöcke im CSV-Format, welche wiederum jeweils eine Größe von ca. 1,5 – 2,5 GB haben. Bezogen werden kann der Datensatz über den Webauftritt von Academic Torrents[12], welche einen Index für wissenschaftliche Daten betreiben.

Sowohl die Tarif- als auch die Reisedaten beinhalten jeweils ca. 14 Millionen Zeilen an Informationen zu einzelnen Taxifahrten, welche einem Taxi eindeutig zugeordnet werden können.

---

[10] www.academictorrents.com/details/107a7d997f331ef4820cf5f7f654516e1704dccf
[11] www.academictorrents.com/details/6c594866904494b06aae51ad97ec7f985059b135
[12] www.academictorrents.com

	A	B	C	D	E	F	G	H	I	J	K
1	medallion	hack_license	vendor_	pickup_datetime	payment_	fare_	surcharge	mta_tax	tip_	tolls_	total_amount
2	3418135604C	B25386A1F259C	VTS	2013-08-30 07:57:00	CSH	41.5	0	0.5	0	0	42
3	6D3B2A7682(	A603A9D5FAA4(	CMT	2013-08-30 23:26:23	CSH	31	0.5	0.5	0	5.33	37.33
4	6D49E494913	3F0BFE90A5D71	CMT	2013-08-30 09:18:10	CSH	5.5	0	0.5	0	0	6
5	4C4A0AFC43:	BA20A20E2CF85	CMT	2013-08-26 23:27:11	CSH	23	0.5	0.5	0	5.33	29.33
6	1258CA1DF5E	8C14DCF69CAA:	CMT	2013-08-29 10:57:56	CSH	14	0	0.5	0	0	14.5
7	3B0E8DC736[	95E1B89BC718F(	CMT	2013-08-27 11:37:45	CSH	31.5	0	0.5	0	5.33	37.33
8	1A575E39B0[	FAE278EFC2171	CMT	2013-08-29 05:41:52	CSH	13	0.5	0.5	0	0	14
9	A30A0C08B13	B56ECCC02B67A	CMT	2013-08-28 19:53:06	CSH	31.5	1	0.5	0	5.33	38.33
10	EF74C4639B1	3251A220F0653	CMT	2013-08-26 01:08:13	CSH	13	0.5	0.5	0	0	14
11	C647516A7B7	80C4140E2FE178	CMT	2013-08-27 21:59:29	CSH	10.5	0.5	0.5	0	0	11.5

Abbildung 3.1 – Ausschnitt der Tarifdaten des Monats August 2013

Während in den Tarifdaten (siehe Abbildung 3.1) ausschließlich Zahlungsinformationen sowie die dazugehörigen Zuschläge, wie z.B. Mautgebühren, zu finden sind, bieten die Reisedaten (siehe Abbildung 3.2) weiterführende Einblicke. So erhält man hier neben der Anzahl der beförderten Personen, der zurückgelegten Strecke wie auch der benötigten Zeit auch Auskunft über den Breiten- und Längengrad des jeweiligen Start- und Zielorts.

	A	B	C	D	E	F	G	H	I	J
1	medallion	hack_license	vendor	rate	store	pickup_datetime	dropoff_datetime	passenger	trip_time	trip_distance
2	3418135604	B25386A1F259	VTS	1		2013-08-30 07:57:00	2013-08-30 08:30:00	5	1980	14.58
3	6D3B2A768:	A603A9D5FAA	CMT	1	N	2013-08-30 23:26:23	2013-08-30 23:46:01	2	1177	11.00
4	6D49E49491	3F0BFE90A5D7	CMT	1	N	2013-08-30 09:18:10	2013-08-30 09:24:08	1	357	.80
5	4C4A0AFC4	BA20A20E2CF8	CMT	1	N	2013-08-26 23:27:11	2013-08-26 23:42:49	4	938	7.70
6	1258CA1DF:	8C14DCF69CA/	CMT	1	N	2013-08-29 10:57:56	2013-08-29 11:19:06	2	1270	2.10
7	3B0E8DC73(	95E1B89BC718	CMT	1	N	2013-08-27 11:37:45	2013-08-27 12:00:58	1	1392	10.90
8	1A575E39B(	FAE278EFC217	CMT	1	N	2013-08-29 05:41:52	2013-08-29 05:55:23	3	811	3.50
9	A30A0C08B	B56ECCC02B67	CMT	1	N	2013-08-28 19:53:06	2013-08-28 20:21:04	1	1678	10.00
10	EF74C4639E	3251A220F065	CMT	1	Y	2013-08-26 01:08:13	2013-08-26 01:19:52	1	697	3.60
11	C647516A7E	80C4140E2FE1	CMT	1	N	2013-08-27 21:59:29	2013-08-27 22:10:07	1	638	2.60

Abbildung 3.2 – Ausschnitt der Reisedaten des Monats August 2013

Es wurde bewusst ein Datensatz gewählt unter dem sich jede Person etwas vorstellen und auch eine Verbindung zu aufbauen kann. Zugleich ist diese Datenmenge für andere Beförderungsunternehmen des Nah- oder Fernverkehrs repräsentativ, zumal sich auch dort ähnliche Daten erfassen lassen. Da sich die Dienstleistungen bzw. Beförderungen von Waren neben dem Straßenverkehr auch auf andere Verkehrsmittel – Schienen-, Flug- sowie Schiffsverkehr – ausweiten lassen, wird eine Vielzahl von Unternehmen angesprochen, welche von dem Einsatz von BI-Software profitieren könnten.

Gerade in dieser Branche herrscht ein ständiger Termin- und Kostendruck. Fahrzeugauslastungen sowie Leerkilometer können anhand der Auswertungen und einer Andersgestaltung der Tourenplanung optimiert werden. Auch die Ursachen für Verspätungen oder plötzlich ansteigender Kosten können besser herauskristallisiert

werden. Ferner besteht generell die Möglichkeit wichtige Informationen schneller weiterzugegeben und so wesentlich flexibler auf verschiedenste Marktveränderungen zu reagieren.

Nichts desto trotz ist diese Datenveröffentlichung auch mit einem skeptischen Auge zu sehen. Wären in diesem Fall die Identifikationsnummern der Taxis nicht verschlüsselt, könnte jede Person z.B. anhand von Fotos mit Zeitstempel beim Einstieg das Fahrziel eines jeden Kunden nachvollziehen. Halbwegs simpel wäre dies bei prominenten Fahrgästen, da sich hier schnell passende Fotos in Suchmaschinen finden lassen und die Identifikationsnummern nahezu aus jedem Blickwinkel ersichtlich sind. Die so einfache Bestimmung der Fahrziele eröffnet Möglichkeiten der Spekulation über deren Erledigungen oder Termine. Da auch die Trinkgelder ersichtlich sind, können sogar Rückschlüsse auf Charaktereigenschaften gebildet werden.

3.2 Datenprüfung

Der von der TLC bereitgestellte Datensatz erfüllt die Qualitätsmerkmale (vgl. Abschnitt 2.2) weitestgehend. Es handelt sich hierbei um eine detailliert aufbereitete Datenmenge, dessen Entstehung nachvollziehbar ist und dessen Sachverhalt ohne jegliche Bewertung wiedergegeben wird. Dementsprechend werden alle Qualitätsmerkmale der intrinsischen Datenqualität erfüllt.

Selbiges gilt auch für die kontextuelle Datenqualität. Zum einen stellt diese ausgiebige Datenmenge ein ganzes Geschäftsjahr lückenlos dar, so dass von einer Erfüllung des Informationsbedarfs ausgegangen werden kann und zum anderen kann sich das Unternehmen einen potenziellen Nutzen von der Analyse und Auswertung dieser Daten erhoffen. Lediglich der Punkt Aktualität wird hier nicht erfüllt. Da jedoch ein vergangenes Jahr analysiert werden soll, ist dieser Punkt weniger relevant.

Darüber hinaus sind die Daten kurz und bündig in getrennt, beschrifteten Spalten dargestellt, wodurch die Verständlichkeit gefördert wird. Die zum Teil anonymisierten Daten beeinträchtigen die Interpretierbarkeit nicht. Ferner konnten auf dem ersten Blick auch keine Widersprüche in sich festgestellt werden, so dass die repräsentationelle Datenqualität ebenfalls erfüllt wird.

Die Zugriffsqualität kann in diesem Fall nicht geprüft werden, da die Daten bereits lokal vorliegen und nicht periodisch in das Data-Warehouse geladen werden.

3.3 Szenarien

Die Analyse dieser Vielzahl an Informationen ermöglicht es, wertvolle Erkenntnisse in den verschiedenen Unternehmensbereichen zu gewinnen und somit Maßnahmen zu ergreifen, welche einen positiven Effekt auf das Unternehmen haben können.

Im Bereich Marketing können durch Auswertung der geografischen Daten die beliebtesten Fahrziele bestimmt werden und so eine Überlegung angestellt werden, ob ggf. eine Bewerbung dieser zu gesonderten Tarifen lohnenswert ist. Hierfür bieten sich auch verschiedene Ansichten wie z.B. die beliebtesten Fahrten zu einer bestimmten Uhrzeit oder an einem bestimmten Tag an.

Ebenso kann der Service optimiert werden, indem die Anzahl der beförderten Fahrgäste untersucht und so festgestellt wird, ob die Kapazitäten der Taxis ausreichend sind oder ggf. eine Einführung von Großraumfahrzeugen eine sinnvolle Alternative darstellt. Darüber hinaus kann auch hier eine Analyse der geografischen Daten stattfinden, indem die Startpunkte betrachtet und so ggf. die Bereitstellungen der Taxis optimiert werden. Dies hätte eine Reduzierung der Wartezeit für die Fahrgäste zur Folge. Auch kann die Nutzung der zur Verfügung gestellten Zahlungsarten ausgewertet und ein Wegfall bzw. eine Hinzufügung weiterer Alternativen in Betracht gezogen werden.

Bei Betrachtung der logistischen Aspekte können die einzelnen Strategien der verschiedenen Fahrer näher in Augenschein genommen werden. So lässt sich erschließen, ob das Warten an beliebten Orten oder das Herumfahren effektiver ist. Die von Erfolg gekrönte Strategie kann so an andere Fahrer vermittelt werden. Ferner können die Fahrzeiten sowie die zurückgelegten Strecken insgesamt sowie individuell untersucht werden. Anhand dieser kann man auch erkennen, wie viele besetzte Taxis unterwegs sind und so die Anzahl je nach Bedarf erhöhen oder vermindern. Überdies lassen sich anhand der zurückgelegten Strecke im Verhältnis zur Fahrzeit auch die Geschwindigkeiten der jeweiligen Taxis bestimmen, so dass man einen Einblick auf die Risikobereitschaft von Bußgeldern erhält. Sofern das Unternehmen diese Kosten übernimmt, können die betroffenen Fahrer entsprechend darauf hingewiesen werden ihr Fahrverhalten zu ändern.

Auch die Personalabteilung kann ihre Vorzüge aus der Analyse ziehen. So können monatlich oder auch jährlich die umsatzstärksten Fahrer bestimmt und zur Steigerung der Mitarbeitermotivation prämiert werden. Obendrein kann eine Bekanntmachung der Uhrzeiten der höchsten Trinkgelder erfolgen. In gleicher Weise kann dies die Motivation steigern und ggf. zur Verbesserung der Kundenfreundlichkeit dienen.

In dem vermutlich wichtigsten Bereich Controlling ist es möglich sämtliche Kosten- und Erlöspunkte wie z.B. die Mautgebühren oder die Fahreinnahmen zu kumulieren und zu veranschaulichen. Alternativ können diese individuell für jedes Taxi dargestellt werden. Um

tiefere Einblicke zu erhalten, lassen sich die Einnahmen u.a. auf einen Monat, einer Woche und sogar auf einen Tag herunterbrechen. Dies ermöglicht z.B. die Feststellung der einnahmenreichsten Woche bzw. des einnahmenreichsten Tags. Weiter ist auch die Bestimmung der zahlungskräftigsten aber auch zahlungsschwächsten Uhrzeit in unterschiedlichen Zeitperioden möglich.

Die nachfolgende Tabelle 3.1 fasst diese Analyseziele noch einmal zusammen.

UNTERNEHMENSBEREICH	ANALYSEZIELE
MARKETING	- Bestimmung der beliebtesten Fahrziele
SERVICE	- Prüfung, ob Fahrzeugkapazitäten ausreichend sind - Prüfung, ob angebotene Zahlungsarten ausreichend oder zum Teil überflüssig sind - Bereitstellung der Taxis optimieren
LOGISTIK	- Ermittlung effektiver Vorgehensweisen zur Gewinnung von Kunden - Bestimmung der bestmöglichen Auslastung - Bußgeldminimierung
PERSONAL	- Bestimmung umsatzstärkster Fahrer - Bestimmung trinkgeldstärkster Zeiten
CONTROLLING	- Kumulierung aller Einnahmen und Veranschaulichung in unterschiedlichen Zeitperioden - Kumulierung aller Ausgaben und Veranschaulichung in unterschiedlichen Zeitperioden - Bestimmung umsatzstärkster bzw. -schwächster Zeiten

Tabelle 3.1 – Überblick Ziele der Datenanalyse

4. Konzeption und Realisierung

Dieses Kapitel beschäftigt sich mit der Umsetzung der gesetzten Analyseziele innerhalb der Pentaho Anwendungen.

4.1 Datenaufbereitung und -integration

Bevor es zu einer Analyse der Daten kommen kann, müssen diese gemäß dem Big Data Life Cycle (vgl. Abschnitt 2.2.1) vorab aufbereitet werden. Dieser Prozess umfasst dabei die Phasen der Datenerfassung, Informationsextraktion sowie Datenintegration.

Pentaho stellt für diese Zwecke die Anwendung Pentaho Data Integration (vgl. Abschnitt 2.4.1) bereit, welche in der Lage ist solch ETL-Prozesse (vgl. Abschnitt 2.1.1.1) umzusetzen. Für die Datenerfassung kann der Anwender zwischen einer Reihe von Möglichkeiten wählen. Dabei können neben allen gängigen Datenbankanbindungen auch andere Optionen wie z.B. Google Analytics, Excel-Tabellen, RSS-Feeds, SAP-Anbindungen, E-Mails, Textdateien in den unterschiedlichsten Formaten, Big Data-Anbindungen oder auch bereits vorbereitete Cubes im XML-Format verwendet werden. In diesem Fall ist die Phase der Datenerfassung aufgrund der Bereitstellung des Datensatzes durch die TLC weitestgehend abgeschlossen.

Durch die Nutzung der „CSV Input“ Option können CSV-Dateien und so auch die lokal vorliegenden Reise- wie auch Tarifdaten der jeweiligen Monate erfasst werden (siehe Anhang 7.1). Die benötigten Informationen lassen sich hier durch das Setzen der in den Dateien enthaltenen Trennzeichen sowie etwaigen Umschließungen extrahieren. Zugleich wird in diesem Schritt die Syntax für die spätere Datenintegration festgelegt, so dass die Daten für diesen Schritt bereits in einem einheitlichen Format vorliegen werden. Zuvor jedoch sind diese noch auf die Datenqualität hin zu prüfen. Die zum Teil schon oberflächlich erfolgte Prüfung dieser im Abschnitt 3.2 gilt es nun zu intensivieren, indem die Datensätze inhaltlich auf Fehler untersucht und von möglich fehlerhaften sowie auch nicht benötigten

Informationen bereinigt werden. So ist es erforderlich die Daten unter individuell festgelegten Bedingungen abfragen und filtern zu können.
Für diese semantischen Transformationen bietet Pentaho Data Integration ebenfalls allerhand Möglichkeiten. So können die Daten beispielsweise sortiert, aggregiert, mit anderen Quellen kombiniert, auf doppelte Angaben geprüft oder durch bestimmte Werte ersetzt werden. Darüber hinaus ist es aber auch möglich diese durch eigens definierte Bedingungen oder auch anhand von SQL- bzw. JSON-Abfragen zu filtern.

Durch Anwendung dieser Filtermöglichkeiten werden die für die spätere Analyse nicht zwingend benötigten Spalten entfernt. Bei den Tarifdaten handelt es sich lediglich um die Spalte „vendor_id“. Die Reisedaten dagegen enthalten neben dieser ebenfalls vorhandenen Spalte auch noch die Spalten „rate_code“ und „store_and_fwd_flag“ (siehe Anhang 7.2). Überdies besteht mittels dieser Option auch die Möglichkeit Spaltennamen zu ändern oder die Reihenfolge dieser zu modifizieren. Solch ein Änderungsbedarf liegt hier jedoch nicht vor.

Aufgrund der zu Beginn gesetzten Syntax für die Spaltentypen können für die inhaltliche Kontrolle eigene Bedingungen formuliert werden, so dass eine Filterung fehlerhafter Datenreihen ermöglicht wird (siehe Anhang 7.3 & 7.4). So wurden für die Längen- und Breitengerade in den Reisedaten beispielsweise Wertebereiche festgelegt. Diese eingegrenzten Bereiche sind darauf zurückzuführen, dass die Angaben in dem sogenannten Dezimalgrad vorliegen. Da dieser entsprechend definiert ist, dürfen die Werte diesen nicht über- bzw. unterschreiten.

Wahlweise kann bestimmt werden, ob die möglichen fehlerhaften Daten in irgendeiner Art und Weise weitere Verwendung finden. So ist es z.B. möglich diese in Textdateien auszulagern oder in gesonderte Datenbanken zu schreiben. Diese könnten so je nach ihrem Stellenwert zu einem späteren Zeitpunkt gesondert untersucht werden, zumal neben den fehlerhaften auch korrekte relevante Daten enthalten sein können. In diesem Fall werden jedoch alle fehlerhaften Datenreihen im weiteren Verlauf ignoriert. Die korrekten dagegen werden direkt in eine vorab angelegte Datenbank geschrieben.

Ein hilfreiches Programmfeature ist hier auch die Preview Funktion. Mit dieser kann zu jedem Zeitpunkt eine gewünschte Anzahl von Datensätzen angezeigt und so überprüft werden, ob die Umsetzung anhand der gesetzten Einstellungen fehlerfrei abläuft.

Die Abbildung 4.1 stellt den final modellierten ETL-Prozess für die Daten des Monats Januar exemplarisch dar, welcher auch mit Hilfe der Sheduling Option periodisch wiederholt werden kann.

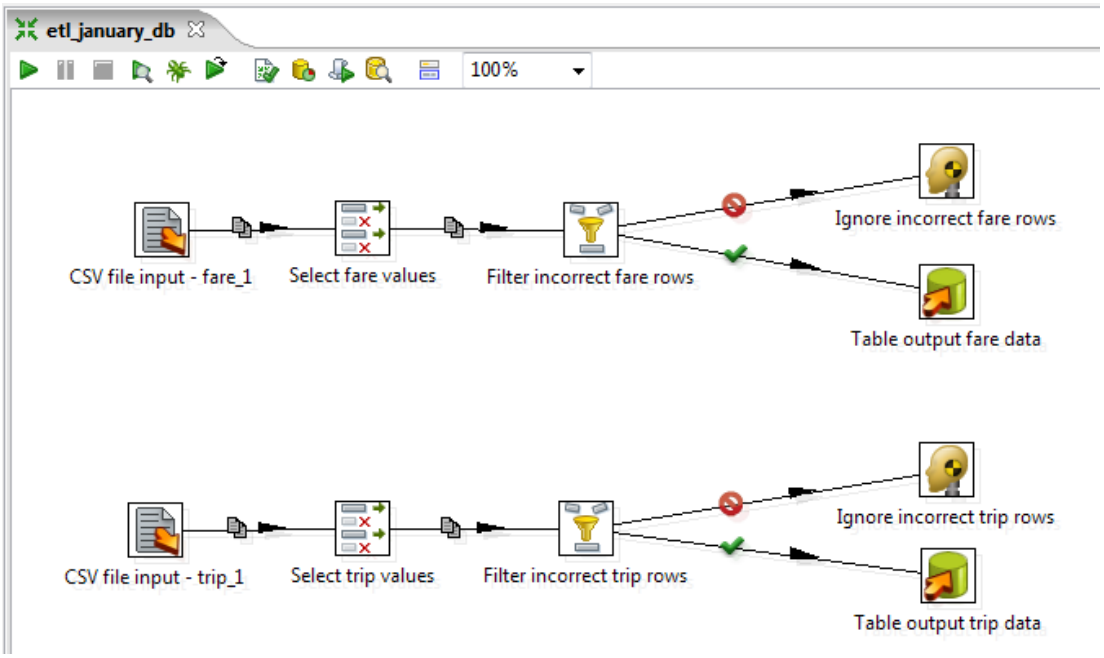


Abbildung 4.1 – ETL-Prozess für den Monat Januar

Folglich konnte der Datensatz durch diesen ETL-Prozess dahingehend bereinigt werden, dass die nun vorliegenden Daten analyserelevant und fehlerfrei sind. Die Größe des Datensatzes wurde dabei von 45,20 GB auf 42,70 GB und die Anzahl der Tabellen von ursprünglich 24 auf zwei reduziert. Die nachfolgende Tabelle 4.1 verschafft einen Überblick der fehlerbehafteten Datenreihen, welche auch bei Ausführung des ETL-Prozesses aufgeführt werden. Hinsichtlich der falschen Tarifdaten ist zu erwähnen, dass diese lediglich im Monat August negative Zahlen oder nur Nullen beinhalteten. Im Gegensatz dazu sind die Fehler in den Reisedaten größtenteils auf die Längen- sowie Breitengerade zurückzuführen. Oftmals waren diese nicht vollständig oder zum Teil mit dem Wert Null besetzt, so dass eine korrekte geografische Bestimmung des Start- bzw. Zielortes nicht mehr möglich war. Vereinzelt war jedoch auch die Angabe der zurückgelegten Strecke oder die benötigte Gesamtzeit fehlerhaft. Abzüglich der fehlerhaften Datenreihen liegen nun 173.172.491 Tarifdatenreihen und 169.342.760 Reisedatenreihen für die Auswertung bereit.

MONAT	EINTRÄGE	FALSCHE TARIFDATEN	FEHLERQUOTE TARIFDATEN	FALSCHE REISEDATEN	FEHLERQUOTE REISEDATEN
JANUAR	14.776.615	0	0,00 %	350.906	2,37 %
FEBRUAR	13.990.176	0	0,00 %	332.683	2,38 %
MÄRZ	15.749.228	0	0,00 %	359.020	2,28 %
APRIL	15.100.468	0	0,00 %	362.772	2,40 %
MAI	15.285.049	0	0,00 %	418.308	2,74 %
JUNI	14.385.456	0	0,00 %	325.502	2,26 %
JULI	13.823.840	0	0,00 %	273.702	1,98 %
AUGUST	12.597.109	7.268	0,06 %	294.979	2,34 %
SEPTEMBER	14.107.693	0	0,00 %	222.839	1,58 %
OKTOBER	15.004.556	0	0,00 %	243.794	1,62 %
NOVEMBER	14.388.451	0	0,00 %	347.209	2,41 %
DEZEMBER	13.971.118	0	0,00 %	305.285	2,19 %
	173.179.759	**7.268**	**0,0042 %**	**3.836.999**	**2,22 %**

Tabelle 4.1 – Fehlerstatistik des Datensatzes der TLC

4.2 Datenanalyse

Die nächste Phase des Big Data Life Cycle ist die Datenanalyse. Hierfür werden die im Abschnitt 3.3 aufgeführten Szenarien herangezogen, indem die vorliegenden, bereinigten Daten Verwendung finden. Ein Großteil der Arbeitsschritte erfolgt hierbei in der Pentaho Business Analytics Plattform CE. Innerhalb dieser können neben einem Administrator beliebig viele Benutzer erstellt und diesen auch verschiedene Rechte eingeräumt werden. Dazu gehören z.B. das Verwalten von Datenquellen oder die Einteilung von Lese-, Schreib-, und Veröffentlichungsrechten.

Innerhalb der Plattform können beliebig viele Datenquellen in Form von CSV-Dateien, SQL-Abfragen oder Datenbankanbindungen für Analysen festgelegt werden. Dem Anwender werden hier zwei Möglichkeiten geboten. Zum einen kann die Anwendung als reines Reportwerkzeug, zum anderen aber auch als Analysewerkzeug verwendet werden. Für Analysen wird in der Regel ein sogenanntes Star-Schema benötigt, in dem die Dimensionen in Form von Tabellen angelegt und die jeweiligen Primär- bzw. Fremdschlüssel dieser in eine Faktentabelle festgelegt werden. Da die in diesem Fall durch den ETL-Prozess erstellten zwei Tabellen (vgl. Abschnitt 4.1) unabhängig voneinander sind, werden diese direkt als Faktentabelle verwendet. Durch diesen Prozess werden programmintern zwei Cubes im XML-Format angelegt, mit denen im weiteren Verlauf gearbeitet werden kann.

Für eine tiefergehende Extraktion von Informationen stellt Pentaho für festgelegte Datenquellen u.a. einen Data Source Model Editor bereit, mit dessen Hilfe Kennzahlen durch Anwendung von einfachen „SUM“, „AVERAGE“, „MINIMUM“, „MAXIMUM“,

„COUNT“ oder „COUNT_DISTINCT“ Aggregationen in Bezug auf eine beliebige Spalte erstellt werden können.

Das Arbeiten mit den erstellten Cubes kann hierbei sowohl mit dem integrierten JPivot Tool aber auch mit einer Reihe verschiedener, durch die Community entwickelter, Plugins geschehen. Diese Plugins können dabei über den eingebundenen Marketplace bezogen werden. Zu Testzwecken wird hier von dieser Möglichkeit auch Gebrauch gemacht und so das Saiku Analytics CE Plugin installiert.

Beide Werkzeuge verfügen über die im Abschnitt 2.1.1.2 erwähnten, gängigen Cube-Operationen wie die Pivotierung, den Drill Down, die Dice Funktion oder auch die Split & Merge Funktion. Auch werden darüber hinaus diverse Visualisierungsmöglichkeiten angeboten. So lassen sich mit Hilfe von Saiku Analytics mögliche Ergebnisse in Form von Balken-, Punkt-, oder Kreisdiagrammen und auch dessen Varianten, Liniendiagrammen, Flächendiagrammen, Tree Maps oder Wasserfalldiagrammen darstellen.
Durch eine Exportfunktion im CSV, XLS und PDF-Format können die Ausgaben der einzelnen Cube-Operationen gespeichert und ggf. verbreitet werden. Bei dem JPivot Tool dagegen sind bzgl. der Visualisierung von Charts kleine Abstriche zu machen. Dennoch sind hier Kreis-, Linien-, Balken- sowie Flächendiagramme und verschiedene Varianten dieser einsetzbar. Darüber hinaus lassen sich hier aber im Vergleich zu Saiku Analytics manuelle Anpassungen, wie z.B. Schriftart und -größe, Breite und Höhe der Chart, Hintergrundfarbe oder auch Informationen zum Titel sowie der Achsenbeschriftung eingeben. Die hier integrierte Exportfunktion unterstützt XLS und PDF-Formate.

Für die dann im Anschluss folgende Interpretationsphase des Big Data Life Cycle gilt es die erzielten Ergebnisse der Datenanalyse für die jeweilige Zielgruppe aufzubereiten und entsprechend bekanntzugeben. Dies kann zum einen innerhalb der Business Analytics Plattform mittels dem Community Dashboard Editor (CDE) oder zum anderen mit dem von Pentaho angebotenen externen Report Designer erfolgen (siehe Abschnitt 2.4.3 & 2.4.4).

Mit Hilfe des CDE lässt sich ein Dashboard entwickeln, welches alle für eine bestimmte Zielgruppe entscheidungsrelevanten Daten vereint und diese innerhalb der Business Analytics Plattform für Entscheidungsträger veranschaulicht darstellt. Darüber hinaus ist dieses auch interaktiv nutzbar und lässt sich jederzeit überarbeiten bzw. erweitern. Der Report Designer dagegen beinhaltet einen gesonderten SQL Query Designer (siehe Anhang 7.5), mit dessen Hilfe entsprechende SQL-Abfragen und somit ebenfalls weitere Auswertungen mit wenig Aufwand erstellt werden können. Diese können auch unmittelbar nach der Erstellung ausgeführt und für eine bestimmte Anzahl an Zeilen als Vorschau dargestellt werden. Das gewünschte Ergebnis kann so vor einer Weiterverarbeitung überprüft werden (siehe Anhang 7.6). Das Resultat einer jeden SQL-Abfrage lässt sich dabei im Report Designer nach Belieben anordnen, in gängige Charts integrieren, mit Labels versehen und auch nach Wunsch formatieren. Fertige Reporte können durch die integrierte

Veröffentlichungsfunktion direkt auf der Business Analytics Plattform veröffentlicht, aber auch lokal z.B. als HTML oder PDF-Dokument gespeichert werden. So kann je nach erteilten Zugriffsrechten jeder registrierte Benutzer der Plattform auf diese zugreifen und von den Ergebnissen profitieren. Die lokalen Dokumente könnten dagegen z.B. als Rundschreiben oder über das Intranet verteilt werden.

Die nachfolgende Tabelle verdeutlicht die in den Analysen zu Trage kommenden Pentaho-Features:

VERWENDUNG	PENTAHO FEATURE
DATENAUFBEREITUNG UND -INTEGRATION	Pentaho Data Integration
DATENQUELLENVERWALTUNG	Data Source Manager inkl. Data Source Model Editor
DATENANALYSE (CUBE-WERKZEUGE)	JPivot Tool Saiku Analytics Plugin CE
BERICHTSWESEN	CDE Pentaho Report Designer

Tabelle 4.2 – Übersicht eingesetzte Pentaho-Features

4.2.1 Marketinganalyse

Wie im Kapitel 3.3 geschildert, sind für den Bereich Marketing die beliebtesten Fahrziele zu bestimmen. Diese Informationen liegen innerhalb des Datensatzes in Längen- sowie Breitengraden vor. Daher wären eine Analyse dieser sowie auch eine Visualisierung mittels einem Geo-Chart wünschenswert. Leider bietet keines der enthaltenen Features solche Charts an.

Abhilfe für diese fehlenden Geo-Charts bietet Saiku Chart Plus, welches eine Erweiterung des Saiku Analytics CE Plugins darstellt und auf das Kartenmaterial von Google zurückgreift. Mit dessen Hilfe können ganze Regionen und auch einzelne Punkte auf einer Weltkarte visualisiert werden. Überdies lassen sich die Charts auch für bestimmte Kontinente oder Länder einschränken. Da jedoch nur maximal 400 Einträge dargestellt werden können und die beliebtesten Ziele nicht innerhalb der Business Analytics Plattform bestimmt werden können, ist dieser Ansatz hier bei Beförderungen im Millionenbereich weniger hilfreich (Google Charts 2015). Nichts desto trotz sind solche Charts z.B. für produzierende Unternehmen und deren Verkaufszahlen ideal geeignet (siehe Anhang 7.7).

Um dennoch die beliebtesten Fahrziele zu bestimmen, wird ausschließlich mit dem Report Designer gearbeitet und mit dessen Hilfe folgende SQL-Abfragen realisiert:

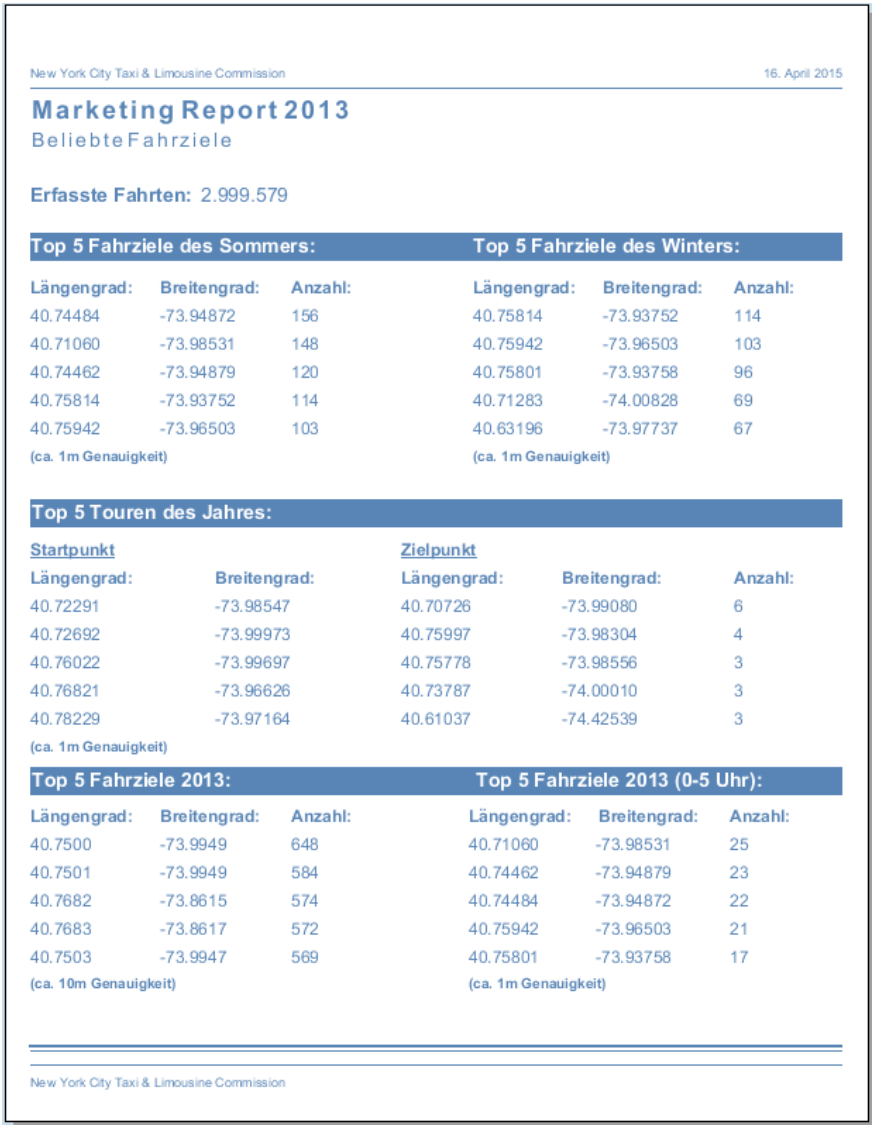
- Anzahl der Gesamtfahrten
- Top 5 Ziele, ca. 10m Genauigkeit
- Top 5 Ziele, ca. 1m Genauigkeit
- Top 5 Ziele, ca. 0,1m Genauigkeit
- Top 5 Touren des Jahres, ca. 1m Genauigkeit
- Top 5 Ziele im Sommer, Frühling, Herbst und Winter, ca. 1m Genauigkeit
- Top 5 Ziele nachts (0-5 Uhr), ca. 1m Genauigkeit

Die Genauigkeit der Ziele ist hierbei von der Anzahl der Nachkommastellen in den vorliegenden Koordinaten abhängig (OpenStreetMap 2015). Die Abbildung 4.2 zeigt eine solche Abfrage exemplarisch für die Zielpunkte mit einer ungefähren Genauigkeit von 0,1m, wofür sechs Nachkommastellen zu berücksichtigen sind.

```
SELECT
    SUBSTRING(`dropoff_latitude`, 1,9) AS `dropoff-latitude`,
    SUBSTRING(`dropoff_longitude`, 1,10) AS `dropoff-longitude`,
    COUNT(*) AS `count`

FROM
    `tripdata`

GROUP BY
    `dropoff-latitude`, `dropoff-longitude`

ORDER BY
    `count` DESC

LIMIT
    5;
```

Abbildung 4.2 – SQL-Abfrage Top 5 Ziele, ca. 0,1m Genauigkeit

Eine Auswahl dieser Abfragen wurde innerhalb des Report Designers platziert, mit Label versehen und entsprechend formatiert (siehe Anhang 7.8). Die Abbildung 4.3 stellt das Ergebnis der Marketinganalyse dar.



New York City Taxi & Limousine Commission 16. April 2015

Marketing Report 2013

Beliebte Fahrziele

Erfasste Fahrten: 2.999.579

Top 5 Fahrziele des Sommers:

Längengrad:	Breitengrad:	Anzahl:
40.74484	-73.94872	156
40.71060	-73.98531	148
40.74462	-73.94879	120
40.75814	-73.93752	114
40.75942	-73.96503	103

(ca. 1m Genauigkeit)

Top 5 Fahrziele des Winters:

Längengrad:	Breitengrad:	Anzahl:
40.75814	-73.93752	114
40.75942	-73.96503	103
40.75801	-73.93758	96
40.71283	-74.00828	69
40.63196	-73.97737	67

(ca. 1m Genauigkeit)

Top 5 Touren des Jahres:

Startpunkt		Zielpunkt		
Längengrad:	Breitengrad:	Längengrad:	Breitengrad:	Anzahl:
40.72291	-73.98547	40.70726	-73.99080	6
40.72692	-73.99973	40.75997	-73.98304	4
40.76022	-73.99697	40.75778	-73.98556	3
40.76821	-73.96626	40.73787	-74.00010	3
40.78229	-73.97164	40.61037	-74.42539	3

(ca. 1m Genauigkeit)

Top 5 Fahrziele 2013:

Längengrad:	Breitengrad:	Anzahl:
40.7500	-73.9949	648
40.7501	-73.9949	584
40.7682	-73.8615	574
40.7683	-73.8617	572
40.7503	-73.9947	569

(ca. 10m Genauigkeit)

Top 5 Fahrziele 2013 (0-5 Uhr):

Längengrad:	Breitengrad:	Anzahl:
40.71060	-73.98531	25
40.74462	-73.94879	23
40.74484	-73.94872	22
40.75942	-73.96503	21
40.75801	-73.93758	17

(ca. 1m Genauigkeit)

New York City Taxi & Limousine Commission

Abbildung 4.3 – Ergebnisdarstellung der Marketinganalyse

4.2.2 Serviceanalyse

Für den Servicebereich sind die Fahrzeugkapazitäten, die Zahlungsarten sowie auch die Startpunkte der Taxis zu untersuchen (vgl. Abschnitt 3.3). Bzgl. der Fahrzeugkapazitäten ist eine Aufschlüsselung der beförderten Personen anhand der erfolgten Beförderungen notwendig. Da derzeit keinerlei Informationen über die Anzahl der Beförderungen vorhanden sind, müssen diese für diesen Zweck generiert werden. Dies geschieht durch

Nutzung des Data Source Model Editors und einer „COUNT“ Funktion (siehe Anhang 7.9). Um die nun nötige Aufschlüsselung zu erhalten, sind für den weiteren Verlauf Slice Funktionen nötig.

Verwendet wird hierfür das Saiku Analytics CE Plugin, indem ausschließlich die Reisedaten als Cube importiert werden. Durch Auswahl der Dimension „Passenger count“ in Kombination mit der zuvor angelegten Kennzahl können die nötigen Slice Funktionen und so auch das erste festgelegte Ziel – Prüfung der Fahrzeugkapazitäten – realisiert werden (siehe Abbildung 4.4).

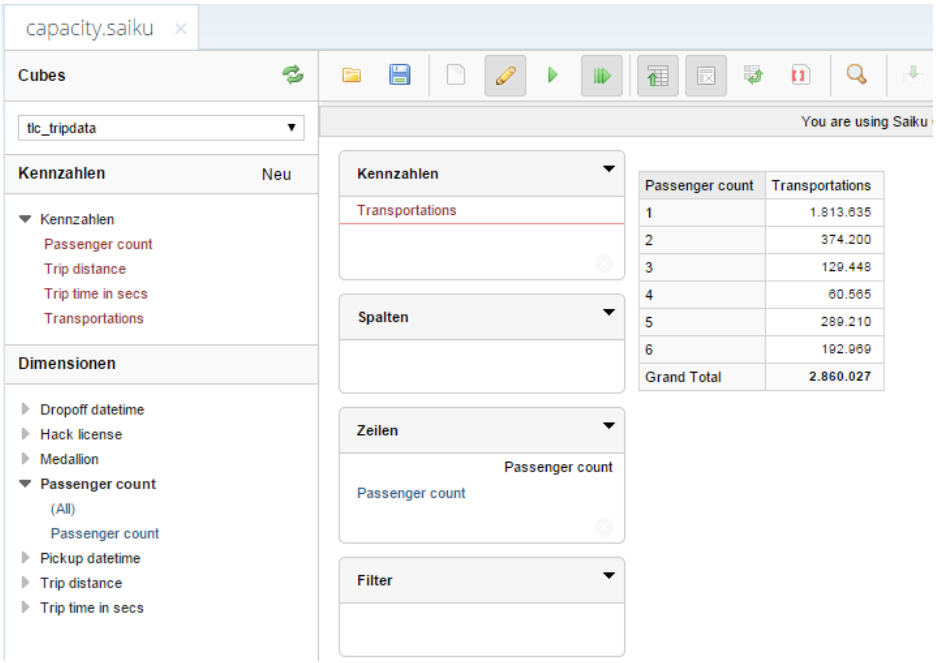


Abbildung 4.4 – Analyse der Fahrzeugkapazitäten

Für die Visualisierung dieses Ergebnisses bietet sich ein Kreis- oder Balkendiagramm an. Der Anwender kann außerdem mit diesen Diagrammen interaktiv agieren. So werden z.B. bei einem Mouseover über einen bestimmten Ausschnitt Layer erstellt, welche die Informationen nochmals detailliert auflisten (siehe Anhang 7.10).

Die Analyse der Zahlungsarten funktioniert nach demselben Prinzip. Jedoch findet zuvor ein Cubewechsel auf die Tarifdaten statt, indem ebenfalls die Kennzahl für die Anzahl der Beförderungen kreiert wird. So muss im weiteren Verlauf lediglich die Dimension „Passenger count“ durch „Payment type“ ersetzt werden (siehe Abbildung 4.5).

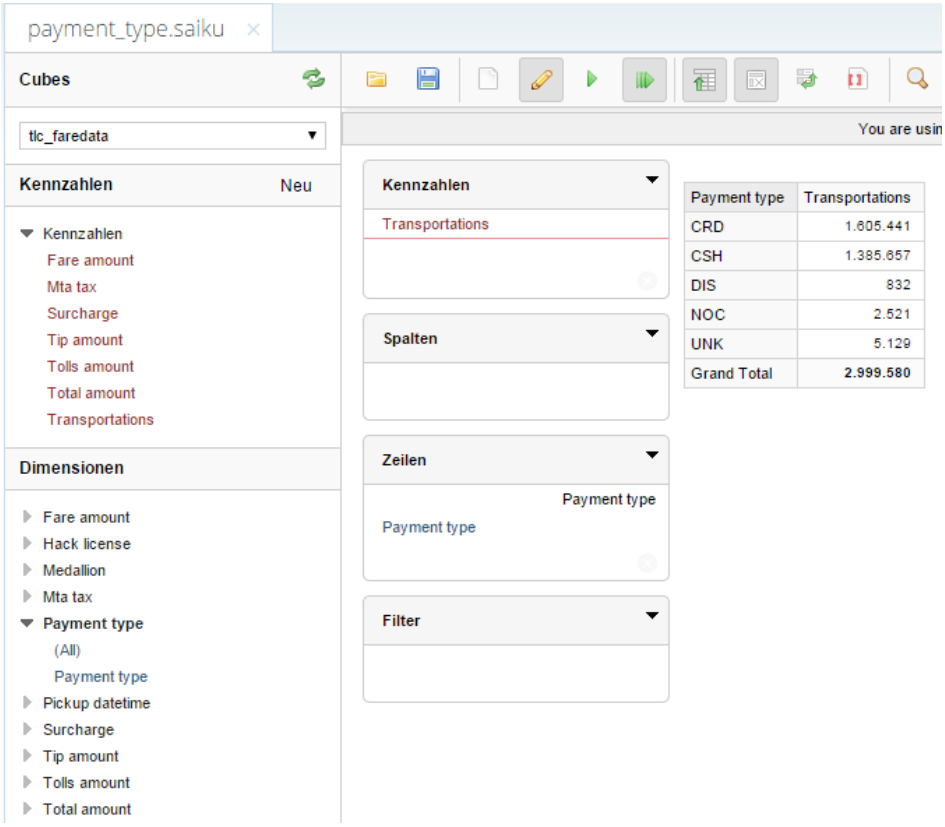


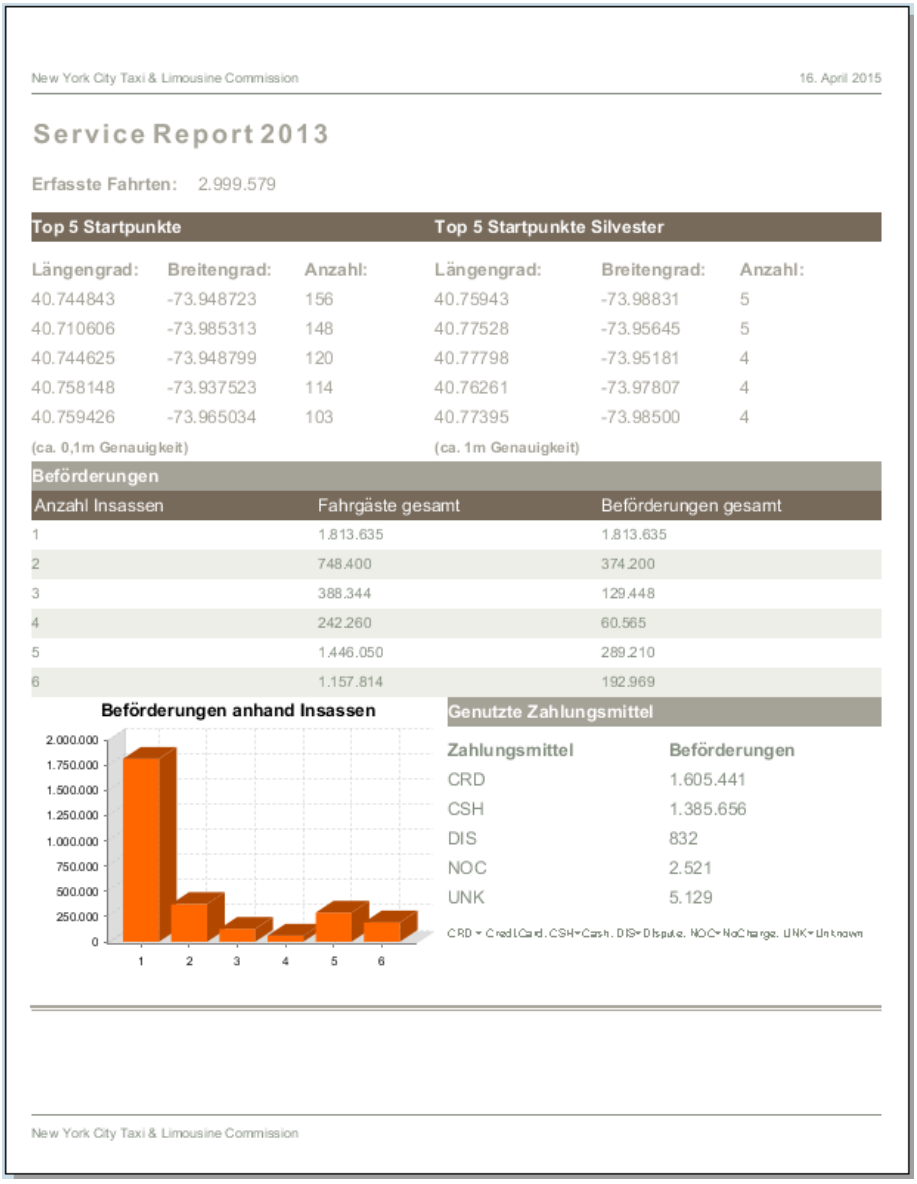
Abbildung 4.5 – Analyse der Zahlungsarten

Die Abkürzungen der angezeigten Zahlungsarten stehen von oben nach unten für Credit Card, Cash, Dispute, No Charge sowie Unknown. Aufgrund der hier unausgewogenen Verhältnisse und der fehlenden Möglichkeit die Skalierung von Charts manuell festzulegen, ist eine Visualisierung mittels dieser nicht zu empfehlen.

Für die Bereitstellung der Taxis und die damit verbundene geografische Auswertung sowie die Ergebnisdarstellung wird auch hier auf den Report Designer zurückgegriffen und folgende SQL-Abfragen erstellt:

- Anzahl der Gesamtfahrten
- Top 5 Startpunkte, ca. 10m Genauigkeit
- Top 5 Startpunkte, ca. 1m Genauigkeit
- Top 5 Startpunkte, ca. 0,1m Genauigkeit
- Top 5 Startpunkte Silvester, ca. 1m Genauigkeit
- Zahlungsmittel anhand Beförderungen
- Anzahl der Insassen anhand der Beförderungen sowie Fahrgästen

Das daraus resultierende Layout (siehe Anhang 7.11) wird in der Abbildung 4.6 in der Ergebnisansicht dargestellt.



New York City Taxi & Limousine Commission 16. April 2015

Service Report 2013

Erfasste Fahrten: 2.999.579

Top 5 Startpunkte			Top 5 Startpunkte Silvester		
Längengrad:	Breitengrad:	Anzahl:	Längengrad:	Breitengrad:	Anzahl:
40.744843	-73.948723	156	40.75943	-73.98831	5
40.710606	-73.985313	148	40.77528	-73.95645	5
40.744625	-73.948799	120	40.77798	-73.95181	4
40.758148	-73.937523	114	40.76261	-73.97807	4
40.759426	-73.965034	103	40.77395	-73.98500	4
(ca. 0,1m Genauigkeit)			(ca. 1m Genauigkeit)		

Beförderungen

Anzahl Insassen	Fahrgäste gesamt	Beförderungen gesamt
1	1.813.635	1.813.635
2	748.400	374.200
3	388.344	129.448
4	242.260	60.565
5	1.446.050	289.210
6	1.157.814	192.969

Beförderungen anhand Insassen

Genutzte Zahlungsmittel

Zahlungsmittel	Beförderungen
CRD	1.605.441
CSH	1.385.656
DIS	832
NOC	2.521
UNK	5.129

CRD = CreditCard, CSH = Cash, DIS = Dispute, NOC = NoCharge, UNK = Unknown

New York City Taxi & Limousine Commission

Abbildung 4.6 – Ergebnisdarstellung der Serviceanalyse

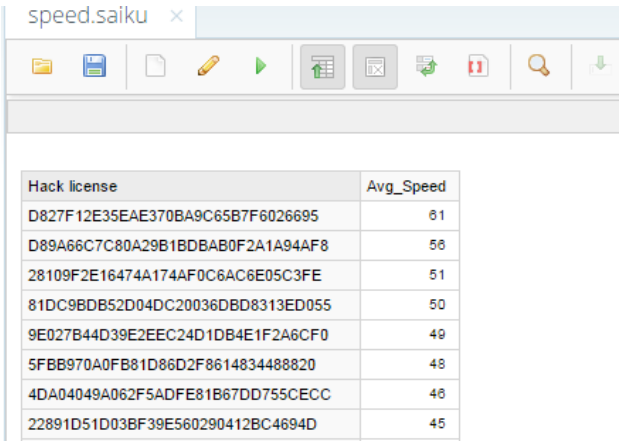
4.2.3 Logistikanalyse

Gemäß den in Abschnitt 3.3 gesetzten Analyseziele für den Bereich Logistik sind hier neben der Bestimmung einer effektiven Vorgehensweise zur Gewinnung von Kunden auch die Auslastung der Taxis sowie die Geschwindigkeiten der Fahrer in Augenschein zu nehmen.

Da für die Bestimmung der Effektivität sowohl Tarif- als auch Reisedaten benötigt werden, ist solch eine Untersuchung ohne größeren Aufwand mittels der Cube-Werkzeuge nicht möglich. Dies ist darauf zurückzuführen, dass zeitgleich immer nur eine Faktentabelle untersucht werden kann. Aus diesem Grund werden die Eckdaten des umsatzstärksten Fahrers, welche Hinweise einer effektiveren Vorgehensweise aufweisen können mit dem Report Designer dargestellt (siehe Anhang 7.12). Auch die Auslastung der Taxis wird über diesen Weg analysiert. Hierfür werden zwei Zeitperioden betrachtet. Zum einen ganztags und zum anderen nur nachts von 0-5 Uhr.

Für die Bestimmung der Geschwindigkeiten jedoch fehlt selbige Information in den Daten. Damit diese dennoch innerhalb der Cube-Werkzeuge oder auch in einer anderen Anwendung ausgewertet werden können, wird ein kleiner ETL-Prozess mit Data Integration erstellt und ausgeführt (siehe Anhang 7.13). Dieser fügt den Reisedaten eine weitere Spalte „Avg_Speed“ hinzu und füllt diese zugleich mit dem entsprechenden Inhalt, indem die zurückgelegte Strecke mit der benötigten Zeit ins Verhältnis gesetzt wird. Im weiteren Schritt werden die Datensätze erneut überprüft und solche in denen eine unglaubwürdige Durchschnittsgeschwindigkeit von über 150 mph erreicht wird entfernt.

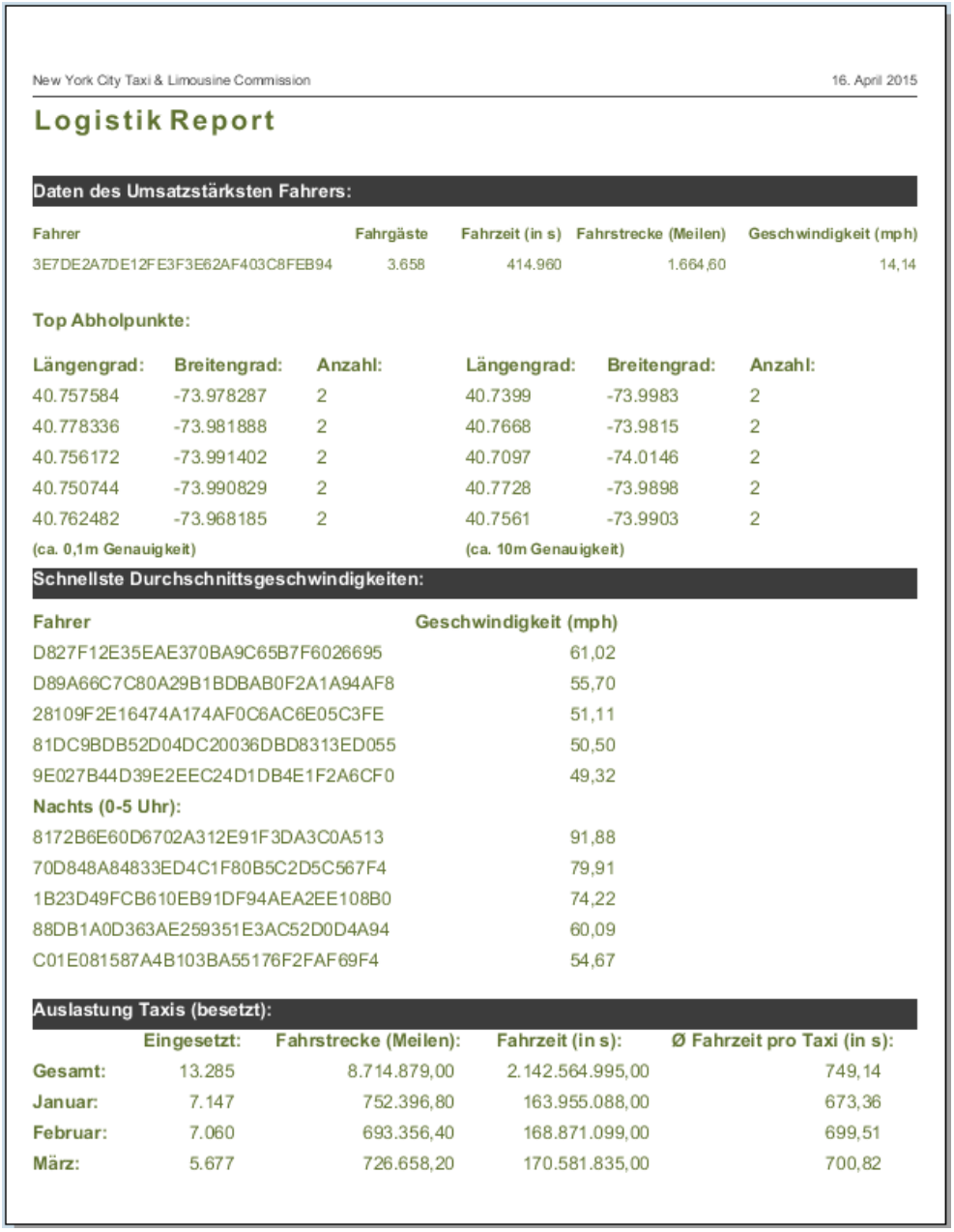
Anschließend können die Geschwindigkeiten der jeweiligen Fahrer mittels Saiku Analytics und Auswahl der Dimensionen „Hack_license“ sowie „Avg_Speed“ dargestellt werden (siehe Abbildung 4.7). Eine Visualisierung mittels einem Chart ist hier bei der Vielzahl an verschiedenen Fahrern nicht angebracht.



Hack license	Avg_Speed
D827F12E35EAE370BA9C65B7F6026695	61
D89A66C7C80A29B1BDBAB0F2A1A94AF8	56
28109F2E16474A174AF0C6AC6E05C3FE	51
81DC9BDB52D04DC20036DBD8313ED055	50
9E027B44D39E2EEC24D1DB4E1F2A6CF0	49
5FBB970A0FB81D86D2F8614834488820	48
4DA04049A062F5ADFE81B67DD755CECC	46
22891D51D03BF39E560290412BC4694D	45

Abbildung 4.7 – Durchschnittsgeschwindigkeiten der Taxifahrer

Die Ergebnisse wurden ebenfalls im Report eingefügt, so dass ein zweiseitiger Bericht entstanden ist (siehe Abbildung 4.8 und Anhang 7.14), welcher in der Business Analytics Plattform zum Abruf bereitgestellt wurde.



New York City Taxi & Limousine Commission 16. April 2015

Logistik Report

Daten des Umsatzstärksten Fahrers:

Fahrer	Fahrgäste	Fahrzeit (in s)	Fahrstrecke (Meilen)	Geschwindigkeit (mph)
3E7DE2A7DE12FE3F3E62AF403C8FEB94	3.658	414.960	1.664,60	14,14

Top Abholpunkte:

Längengrad:	Breitengrad:	Anzahl:	Längengrad:	Breitengrad:	Anzahl:
40.757584	-73.978287	2	40.7399	-73.9983	2
40.778336	-73.981888	2	40.7668	-73.9815	2
40.756172	-73.991402	2	40.7097	-74.0146	2
40.750744	-73.990829	2	40.7728	-73.9898	2
40.762482	-73.968185	2	40.7561	-73.9903	2
(ca. 0,1m Genauigkeit)			(ca. 10m Genauigkeit)		

Schnellste Durchschnittsgeschwindigkeiten:

Fahrer	Geschwindigkeit (mph)
D827F12E35EAE370BA9C65B7F6026695	61,02
D89A66C7C80A29B1BDBAB0F2A1A94AF8	55,70
28109F2E16474A174AF0C6AC6E05C3FE	51,11
81DC9BDB52D04DC20036DBD8313ED055	50,50
9E027B44D39E2EEC24D1DB4E1F2A6CF0	49,32
Nachts (0-5 Uhr):	
8172B6E60D6702A312E91F3DA3C0A513	91,88
70D848A84833ED4C1F80B5C2D5C567F4	79,91
1B23D49FCB610EB91DF94AEA2EE108B0	74,22
88DB1A0D363AE259351E3AC52D0D4A94	60,09
C01E081587A4B103BA55176F2FAF69F4	54,67

Auslastung Taxis (besetzt):

	Eingesetzt:	Fahrstrecke (Meilen):	Fahrzeit (in s):	Ø Fahrzeit pro Taxi (in s):
Gesamt:	13.285	8.714.879,00	2.142.564.995,00	749,14
Januar:	7.147	752.396,80	163.955.088,00	673,36
Februar:	7.060	693.356,40	168.871.099,00	699,51
März:	5.677	726.658,20	170.581.835,00	700,82

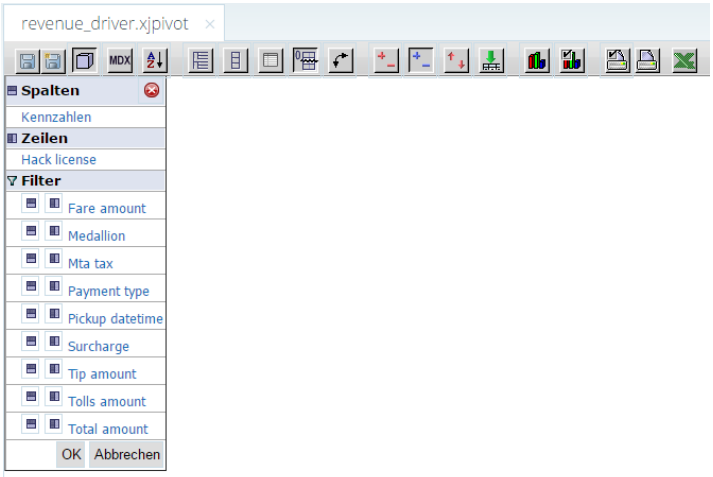
Abbildung 4.8 – Ergebnisdarstellung der Logistikanalyse (Seite 1/2)

4.2.4 Personalanalyse

Die im Abschnitt 3.3 dargestellten Ziele der Personalanalyse beschäftigen sich mit der Bestimmung des umsatzstärksten Fahrers sowie den trinkgeldstärksten Zeiten. Hinsichtlich der Umsatzanalyse ist eine Aufschlüsselung der jeweiligen Fahrer anhand der Einnahmen notwendig. Dabei ist es hier möglich bei der Einnahmenbestimmung zwischen zwei Arten zu unterscheiden. Zum einen können die reinen Fahrpreiseinnahmen und zum anderen die Gesamteinnahmen, welche auch mögliche Mautbeträge, Steuern sowie Trinkgeld enthalten, herangezogen werden. Um hier beide Möglichkeiten zu realisieren, müssen diese fehlenden Kennzahlen ergänzt werden können. Darüber hinaus ist auch interessant zu wissen, wie viele Fahrten der jeweilige Fahrer für seine Einnahmen benötigt hat.

Bezüglich der Analyse des Trinkgelds sind ebenfalls verschiedenste Varianten denkbar. So lassen sich diese z.B. den jeweiligen Fahrern zuweisen, um so evtl. Rückschlüsse auf deren Auftreten gegenüber dem Fahrgast zu erhalten. Alternativ können diese aber auch anhand der Zahlungsarten gruppiert werden. Da es hier primär um den zeitlichen Faktor geht, ist z.B. das durchschnittliche Trinkgeld je Stunde interessant. Eine Aufschlüsselung nach Monaten wäre ebenfalls denkbar. Um dieses zu realisieren, ist es notwendig die Dimension Zeit beliebig zu splitten.

Für die Umsatzanalyse wird hier auf das integrierte JPivot Tool zurückgegriffen. So können im Cube der Tarifdaten die fehlenden Kennzahlen mit dem Data Source Model Editor durch einer „SUM“ Aggregation problemlos gebildet und darüber hinaus auch formatiert werden. Durch einer Filterung der nicht benötigten Dimensionen sowie Auswahl der Kennzahl für die Fahrpreiseinnahmen kann letztendlich das gewünschte Ziel realisiert werden (siehe Abbildung 4.9).



Hack license	Kennzahlen: Fare amount	Kennzahlen: Transportations
All Hack licenses	$ 37.218.467,90	2.999.580
3E7DE2A7DE12FE3F3E62AF403C8FEB94	$ 7.415,50	690
4FBAD760692D8C2B7A2ACE057B3B5F1B	$ 7.312,00	646
75513D3470577109C3C0A5AE755C88AB	$ 7.076,50	548
CAFFA06F9610751C8B17EE3D6DDEA59E	$ 7.005,00	553
C8D349DC61BCD170E1122B6503F8EE78	$ 6.923,00	554
9D5F4C4B01DAC9C86B17751BD8D23421	$ 6.865,00	553
030C2C1ED4F1CEA429F4DA3DAE4F8B62	$ 6.859,80	548
946438FB059349E669781CE23362A176	$ 6.740,00	585
0238047A5834FF6524C69E811F873D5A	$ 6.625,30	528
518BCE0D10BDD35A6AF18DB1B87318BC	$ 6.608,00	556
9C05EA894FF67FAF656CE3604CCD9FBF	$ 6.598,50	504
07570576F0B71F31372110F17DFDB95F	$ 6.586,00	525
B3EF81B3356EBD377C83E11B1085F5D4	$ 6.557,00	596
E87E82B69AD6B0098934FA7EFE15841D	$ 6.529,00	437
91B5E1229B2D1BAF9B3E47F950809 52D	$ 6.516,50	380
C44ECD760F3B383AC1D9593D343F633D	$ 6.468,50	393

Abbildung 4.9 – Analyse der umsatzstärksten Fahrer

Entsprechend wird dies auch für die Gesamteinnahmen umgesetzt (siehe Anhang 7.15). Erwähnenswert ist hier, dass JPivot lediglich eine auf- oder absteigende Sortierung der Ergebnismenge anbietet. Saiku Analytics dagegen bietet noch weitere Optionen, wie z.B. eine vielfältige Filtermöglichkeit oder auch eine begrenzte Ansicht der Datensätze (siehe Anhang 7.16). Eine mögliche Visualisierung dieser Ergebnisse könnte innerhalb von Saiku Analytics mit einem Balkendiagramm vollzogen werden. Aufgrund der fehlenden Möglichkeit bei JPivot die Ausgabe einzuschränken, empfiehlt sich eine dortige Visualisierung nicht.

Für die Analyse der Trinkgelder wird hier auf das Saiku Analytics Plugin zurückgegriffen. Zu Grunde liegt auch hier der Cube der Tarifdaten. Die benötigte Kennzahl kann ebenfalls mit dem Data Source Model Editor und einer „SUM“ Aggregation erstellt werden. Um einen höheren Informationsgehalt zu erhalten, können die in den Analysen zuvor erstellten

Kennzahlen mit einbezogen werden. So werden diese hier anhand der Zahlungsarten gruppiert (siehe Abbildung 4.10).

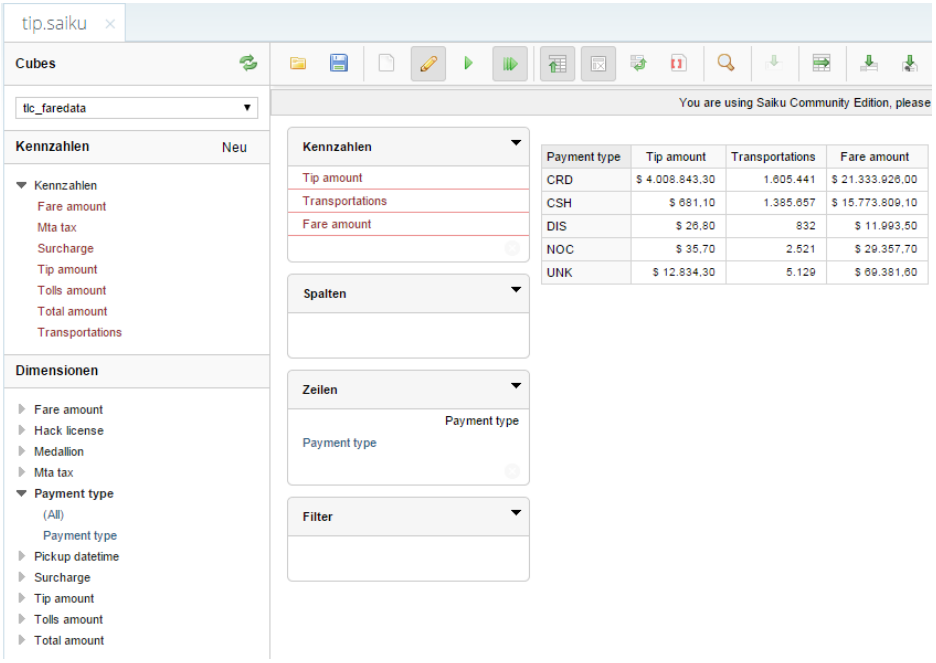


Abbildung 4.10 – Analyse der Trinkgelder gruppiert in Zahlungsarten

Die gewünschten Analysen der Trinkgelder mit einem zeitlichen Bezug können mit dem vorliegenden Cube weder mit dem JPivot Tool noch mit dem Saiku Analytics Plugin umgesetzt werden. Dies ist darauf zurückzuführen, dass die zeitlichen Informationen in der Dimension „Pickup datetime“ nicht weiter aufgeschlüsselt werden können und sich im Data Source Model Editor auch keine hilfreichen Kennzahlen hierfür bilden lassen. Um diese Analysen dennoch umzusetzen sind verschiedene Ansätze möglich. Zum einen kann eine Aufschlüsselung der Zeitinformationen bereits im ETL-Prozess erfolgen. Dadurch stünden nun weitere Dimensionen zur Auswahl bereit, die eine Zuordnung ermöglichen würden. Zum anderen bietet Pentaho das Tool Schema Workbench an. Mit dessen Hilfe kann die dem Cube zugrunde liegende XML Datei bearbeitet werden. So lassen sich hier neben der Bearbeitungsmöglichkeiten für die Hierarchien auch umfangreichere Kennzahlen erstellen, die z.B. auf Abfragen mittels SQL zurückgreifen. Eine dritte und hier weiterführend verwendete Möglichkeit ist ebenfalls der Report Designer. Für die hier festgelegten Ziele wurden mittels diesem Designer folgende Abfragen realisiert:

- Trinkgeld gesamt
- Gesamtfahrten
- Gesamteinnahmen

- Die Top 5 Fahrer eines jeden Monats sowie des kompletten Jahres
- Anzahl eingesetzter Fahrer
- Anzahl eingesetzter Taxis
- Einnahmen pro Monat
- Trinkgeld in % kleiner gleich 100% mit Anzahl der zugehörigen Fahrten
- Trinkgeld in % größer als 100% mit Anzahl der zugehörigen Fahrten
- Fahrpreiseinnahmen
- Fahrten ohne Trinkgeld
- Durchschnittliches Trinkgeld pro Fahrt je Stunde
- Durchschnittliches Trinkgeld pro Fahrt je Monat

Ein Teil der Ergebnisse, welche aus diesen Abfragen resultieren, wurden auch hier im Report Designer platziert und entsprechend formatiert (siehe Anhang 7.17). Die folgende Abbildung 4.11 zeigt den Report in der Vorschauansicht. Dieser kann je nach Wunsch z.B. als PDF-Datei gespeichert und per Mail verteilt werden. Ein direkter Export als HTML-Datei ist ebenfalls möglich. So sind z.B. eine Integration und dadurch ein direkter Abruf im Intranet denkbar.

Abbildung 4.11 – Ergebnisdarstellung der Personalanalyse

4.2.5 Controllinganalyse

Der Bereich Controlling ist insbesondere für die Führungsebene von entscheidender Bedeutung. Die dortigen Bewegungen müssen ständig im Auge behalten werden, um bei etwaigen falschen Entwicklungen schnell handeln zu können. Aus diesem Grund wäre eine Oberfläche, welche die wichtigsten Daten vereint wünschenswert. Führungskräfte erhalten so jederzeit einen Überblick über gewünschte Zeiträume. Daher wird für die im Abschnitt 3.3 aufgeführten Analyseziele des Controllings der integrierte CDE verwendet. So können sämtliche Einnahmen und Ausgaben in unterschiedlichen Zeitperioden sowie auch die umsatzstärksten Zeiten übersichtlich veranschaulicht werden. Die Abbildung 4.12 stellt einen ersten möglichen Entwurf dieser Übersicht dar. Solch ein Dashboard lässt sich zudem jederzeit erweitern und bietet so z.B. auch Möglichkeiten laufende bzw. noch folgende Geschäftsjahre mit vergangenen zu vergleichen.

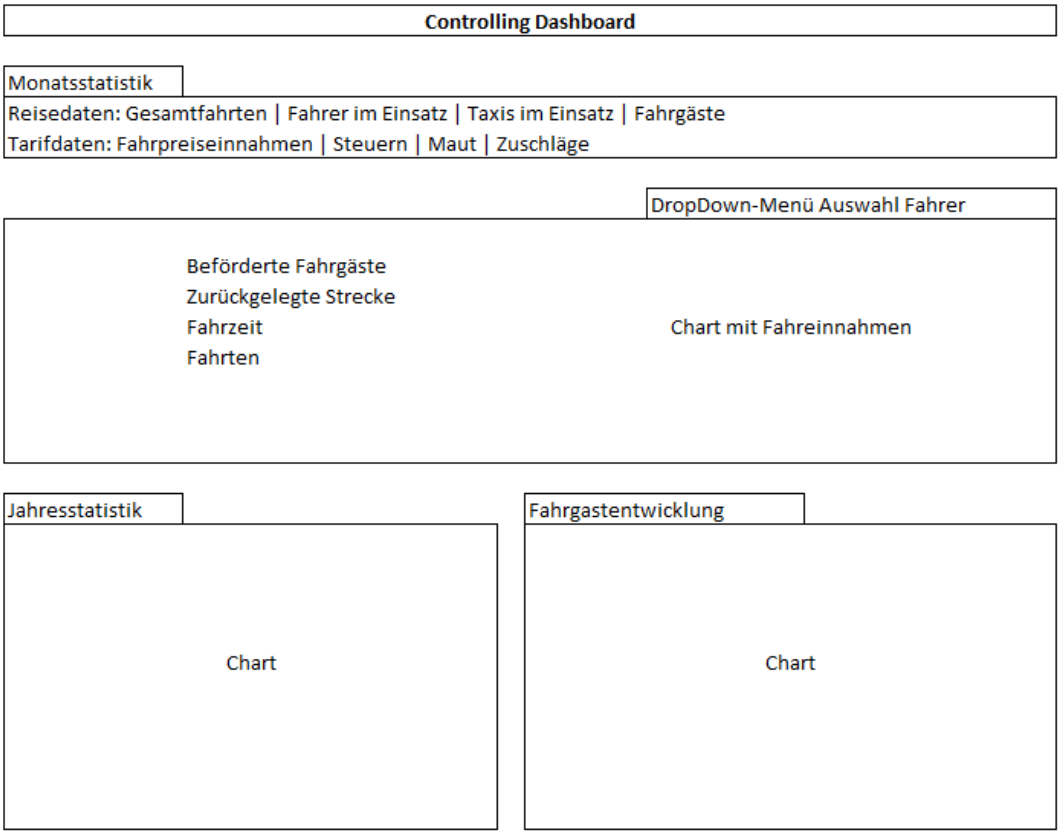


Abbildung 4.12 – Dashboard Layout Entwurf für die Controllinganalyse

Gemäß dem erstellten Entwurf wird innerhalb vom CDE ein Layout angelegt (siehe Anhang 7.18). Dieses ist mit einer Tabelle zu vergleichen, indem je nach Bedarf Zeilen und Spalten hinzugefügt werden können. Auf Wunsch lässt sich dieses auch speziell für mobile

Anwendungen anpassen. Grundlage dieses Layouts ist die Programmiersprache HTML. Sollten einem die vorliegenden Designs nicht zusagen, so kann entweder per JavaScript oder CSS ein eigenes erstellt werden. In diesem Fall wird zu Testzwecken eins mittels CSS entworfen (siehe Anhang 7.19).

Innerhalb dieses Layouts werden anschließend die benötigten Komponenten durch eine Verlinkung platziert. Der CDE bietet hier eine Vielzahl von Möglichkeiten an. So können neben den gängigen Charts z.B. auch ganze Tabellen, Auswahlmenüs, Buttons oder Navigationsleisten integriert werden. In diesem Fall werden neben einem Dropdown-Auswahlmenü für die Fahrer auch 3 Charts sowie 3 Tabellen benötigt (siehe Anhang 7.20).

Damit innerhalb der Komponenten auch Daten dargestellt werden können, benötigen diese wiederum eine Datenquelle. Für gewöhnlich werden hierfür SQL oder MDX Abfragen genutzt. So werden die hier benötigten Daten mit Hilfe von SQL abgegriffen und mit den jeweiligen Komponenten verknüpft. Mittels der SQL-Abfrage in Abbildung 4.13 werden z.B. die Einnahmen eines bestimmten Fahrers, welcher vorab durch ein Drop Down Menü ausgewählt wird, berechnet.

```
SELECT
    MONTH(`faredata`.`pickup_datetime`) AS `Monat`,
    ROUND(SUM(`faredata`.`total_amount`),2) AS `Einnahmen gesamt pro Monat`

FROM
    `faredata`

WHERE
    `hack_license`=${driver}

GROUP BY
    `Monat`
```

Abbildung 4.13 – SQL-Abfrage der Umsätze eines bestimmten Fahrers

Das endgültige Ergebnis des Dashboards zeigt die Abbildung 4.14. Die Fahrerstatistik aktualisiert sich bei Auswahl eines neuen Fahrers automatisch (siehe Anhang 7.21).

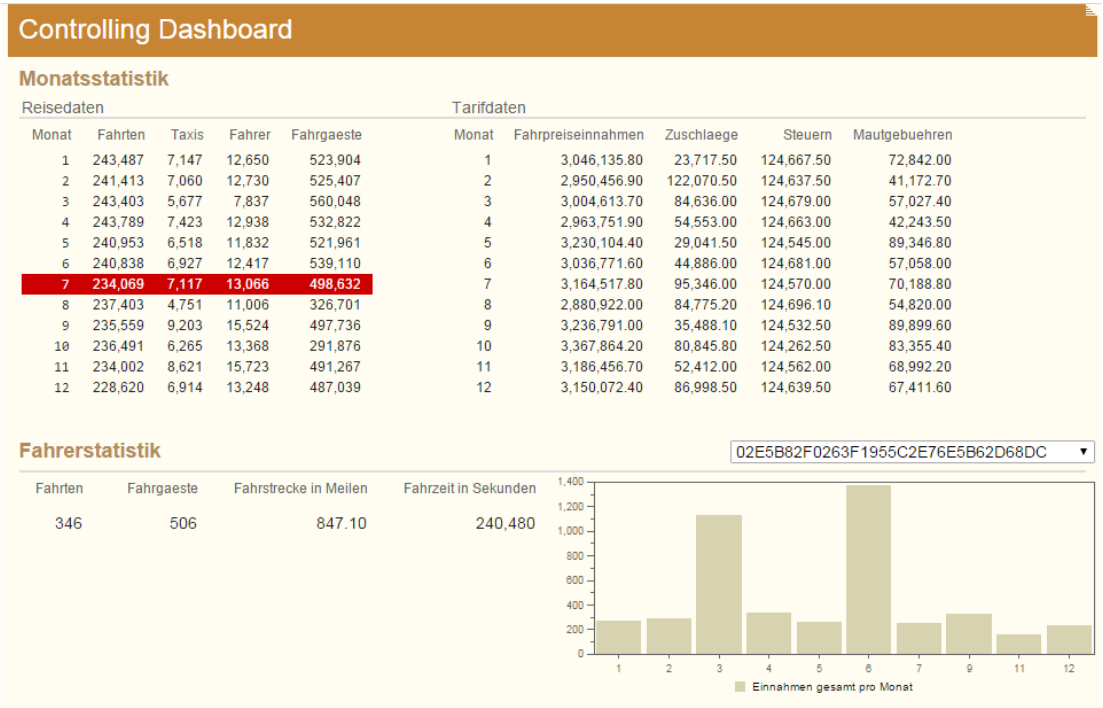


Abbildung 4.14 – Controlling Dashboard

4.3 Vorbereitende Maßnahmen

Für die Realisierung des Vorhabens wurde ein eigener Rechner mit einem lokalen Server sowie folgenden Eigenschaften verwendet:

CPU:	AMD Phenom™ II X6 1100T Processor 3.31 GHz
Mainboard:	ASUS Crosshair IV Formula 890FX
Grafikkarte:	NVidia GeForce GTX 580
Arbeitsspeicher:	8 GB DDR3 SDRam 2000 MHz
Festplatte:	Crucial RealSSD C300 CTFDDAC128MAG - 128 GB
Betriebssystem:	Windows 7 64 Bit Professional Edition
Server:	Apache 2.4.10
	MySQL 5.6.21
	PHP 5.6.3
	phpMyAdmin 4.2.11
Software:	Java 7
	Google Chrome 41.0.2272.118

Pentaho Business Analytics Platform CE 5.2.0.0-209
Pentaho Data Integration CE 5.2.0.0-209
Pentaho Report Designer CE 5.2.0.0-209
Pentaho Marketplace CE 5.2.0.0-209
Plugin: Saiku Analytics CE 3.0.9.1 für die Business Analytics Plattform
Plugin: Saiku Chart Plus, Erweiterung von Saiku Analytics CE 3.0.9.1

Da Pentaho derzeit nur mit Java 7 kompatibel ist, war vorab eine Zurückstufung von Java 8 auf 7 notwendig. Die Umgebungsvariable für „JAVA_HOME“ musste daraufhin ebenfalls aktualisiert werden. Um eine Verbindung zur Datenbank aufbauen zu können, musste außerdem der nicht integrierte MySQL Treiber den jeweiligen Programmverzeichnissen hinzugefügt werden.

Generell können hier auch bekannte Big Data Lösungen wie z.B. Hadoop oder MongoDB zum Einsatz kommen. Beide Möglichkeiten werden auch von Pentaho unterstützt. Anzumerken ist jedoch, dass Pentaho für MongoDB derzeit keinerlei Datenbankanbindungen außerhalb von Pentaho Data Integration oder dem Report Designer anbietet. Folglich lässt sich kein Cube bzw. mit dessen Hilfe Visualisierungen von Dimensionen mittels Kennzahlen erstellen. Dennoch ist die Erstellung eines Dashboards möglich. Die für die Komponenten benötigten Datenquellen müssen hierbei einen ETL-Prozess darstellen. Diese müssen jeweils vorab in Data Integration erstellt und abgespeichert werden. Die jeweils gespeicherte Datei kann daraufhin als Datenquelle genutzt werden.
Da diese Big Data Konzepte jedoch speziell für Rechennetzwerke ausgelegt sind, welche auf dieser Basis ein Sharding bzw. Clustering ermöglichen und der Versuch solch ein Netzwerk virtuell umzusetzen aus Performanzgründen fehlschlug, finden diese Ansätze hier keine Anwendung. Stattdessen wird auf eine MySQL Datenbank zurückgegriffen. Darüber hinaus wurden, ebenfalls aus Performanzgründen, die für die Analyse bereitstehenden Datensätze – gemäß dem Abschnitt 4.1 – auf jeweils ca. drei Millionen reduziert.

5. Bewertung

Dieses Kapitel befasst sich mit der abschließenden Bewertung des Pentaho Open Source Business Intelligence Systems.

5.1 Bewertungsweise

Gemäß Oppermann und Reiterer sind bei einer Evaluation von EDV-Systemen die in der Abbildung 5.1 dargestellten Elemente sowie deren Beziehungen zueinander zu berücksichtigen.

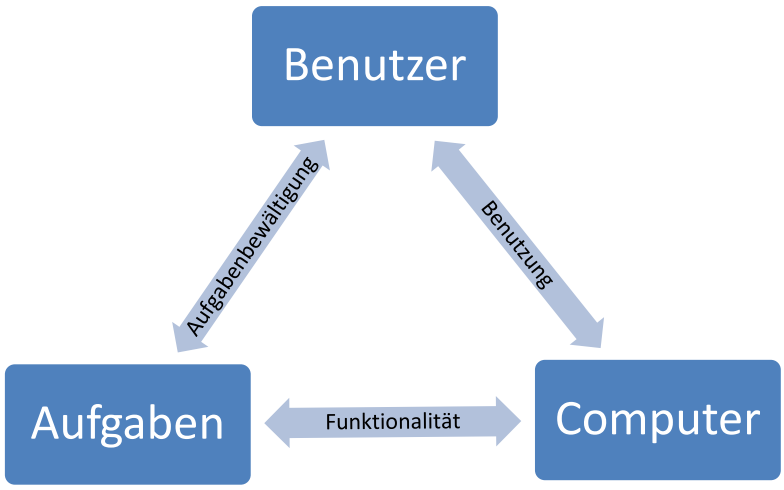


Abbildung 5.1 – Elemente & Beziehungen einer Software-Evaluation anhand (Oppermann und Reiterer 1994)

In der Beziehung zwischen Aufgaben und Benutzer geht es dabei um die Frage, inwieweit der Benutzer in der Lage ist, die Aufgaben zu erfüllen. Hinsichtlich EDV-Systeme wird insbesondere untersucht, ob die Aufgabenbewältigung vom System unterstützt oder gar

behindert wird. Die Verbindung zwischen Aufgaben und Computer dagegen klärt das Ausmaß der Unterstützung der Aufgaben durch das EDV-System. Dabei ist zu untersuchen, ob dieses die Aufgaben zufrieden stellend abbilden kann oder diese verkompliziert. Gleichzeitig liegt das Augenmerk auch auf die Möglichkeit der Umgestaltung bestimmter Aspekte der Arbeitsaufgabe, wie z.B. das Erweitern durch Hinzufügen neuer Funktionalität. Die dritte Beziehung zwischen Benutzer und Computer berücksichtigt inwiefern die vorhandene bzw. nicht vorhandene Funktionalität die Qualität der Benutzung beeinflusst. In Bezug darauf liegt hier der Interaktionsaufwand zwischen dem Benutzer und dem EDV-System im Vordergrund. Da diese Schnittstelle ein zentraler Bewertungsgegenstand von Software-Evaluationen ist, kommt diesem Punkt eine besondere Aufmerksamkeit zu (Oppermann und Reiterer 1994).

In der europäischen Norm ISO 9241-110 sind sieben Kriterien zu finden, welche sich speziell auf die Gestaltung und Bewertung von Benutzungsschnittstellen interaktiver Systeme beziehen (Schneider 2008):

- Aufgabenangemessenheit
    - „Ein interaktives System ist aufgabenangemessen, wenn es den Benutzer unterstützt, seine Arbeitsaufgabe zu erledigen, d. h., wenn Funktionalität und Dialog auf den charakteristischen Eigenschaften der Arbeitsaufgabe basieren, anstatt auf der zur Aufgabenerledigung eingesetzten Technologie.“

- Selbstbeschreibungsfähigkeit
    - „Ein Dialog ist in dem Maße selbstbeschreibungsfähig, in dem für den Benutzer zu jeder Zeit offensichtlich ist, in welchem Dialog, an welcher Stelle im Dialog er sich befindet, welche Handlungen unternommen werden können und wie diese ausgeführt werden können.“

- Lernförderlichkeit
    - „Ein Dialog ist lernförderlich, wenn er den Benutzer beim Erlernen der Nutzung des interaktiven Systems unterstützt und anleitet.“

- Steuerbarkeit
    - „Ein Dialog ist steuerbar, wenn der Benutzer in der Lage ist, den Dialogablauf zu starten sowie seine Richtung und Geschwindigkeit zu beeinflussen, bis das Ziel erreicht ist.“

- Erwartungskonformität
    - „Ein Dialog ist erwartungskonform, wenn er den aus dem Nutzungskontext heraus vorhersehbaren Benutzerbelangen sowie allgemein anerkannten Konventionen entspricht.“

- Individualisierbarkeit
    - „Ein Dialog ist individualisierbar, wenn Benutzer die Mensch-System-Interaktion und die Darstellung von Informationen ändern können, um diese an ihre individuellen Fähigkeiten und Bedürfnisse anzupassen."

- Fehlertoleranz
    - „Ein Dialog ist fehlertolerant, wenn das beabsichtigte Arbeitsergebnis trotz erkennbar fehlerhafter Eingaben entweder mit keinem oder mit minimalem Korrekturaufwand seitens des Benutzers erreicht werden kann."

Um eine bessere Messbarkeit dieser Kriterien zu erlangen wurden für diese gemäß der Tabelle 5.1 weitere Unterkriterien aufgestellt. Aufgrund des modularen Aufbaus von Pentaho (vgl. Abschnitt 2.4), werden die angewandten Features getrennt voneinander auf diese Kriterien hin untersucht und abschließend als Gesamtwertung in einer Nutzwertanalyse dargestellt. Innerhalb dieser ist den einzelnen Features je nach Beitrag zum Gesamtergebnis ein Wichtigkeitsgrad zugeteilt. Die einzelnen Kriterien erhalten dabei ebenfalls eine Gewichtung von bis zu 100%. Anschließend werden je nach Grad der Erfüllung Punkte verteilt und mittels einer Berechnungsformel ein Wert für das jeweilige Feature sowie Kriterium errechnet. Die Gesamtpunkte der einzelnen Anwendungen werden dann kumuliert, so dass eine Gesamtpunktzahl für die BI Suite zustande kommt.

AUFGABENANGEMESSENHEIT	- Werden unnötige Arbeitsschritte vermieden? - Bleiben dem Benutzer überflüssige Informationsanzeigen oder Hilfestellungen erspart? - Können Arbeitsschritte gespeichert werden? - Erfolgt eine aufgabenorientierte Informationsein- und –ausgabe?
SELBSTBESCHREIBUNGSFÄHIGKEIT	- Werden Zustandsänderungen des Dialogsystems angezeigt (z.B. Eingabeerwartung oder Wartezeit)? - Wird das Heranziehen von zusätzlichen externen Informationen für das Verständnis der Anwendung vermieden? - Schützt die Anwendung den Benutzer vor folgenschweren Handlungen (z.B. löschen oder überschreiben von Daten)?
LERNFÖRDERLICHKEIT	- Erfolgen die Bedienschritte nach einem klar nachvollziehbaren Prinzip? - Werden relevante Lernstrategien („Learning-by-doing") unterstützt? - Existieren für komplexe Sachverhalte Tutorials oder andere

	Lernelemente? - Existiert ein durchgängiges Konzept bei der Strukturierung von Dialogen?
STEUERBARKEIT	- Lassen sich Bedienschritte mehrstufiger Eingabeprozesse rückgängig machen (Undo-Funktion)? - Kann der Benutzer Abläufe unterbrechen und zu einem anderen Zeitpunkt an gleicher Stelle fortführen? - Kann der Benutzer die Informationsmenge und Ausgabeart frei bestimmen?
ERWARTUNGSKONFORMITÄT	- Wird das WYSIWYG-Prinzip (What You See Is What You Get) beachtet? - Verfügen ähnliche Arbeitsabläufe auch über ähnliche Dialogverläufe? - Entsprechen die Bedien- und Steuerfunktionen den zu erwartenden Vorerfahrungen der Benutzer?
INDIVIDUALISIERBARKEIT	- Kann die Sprache der Anwendung angepasst werden? - Existieren alternative Darstellungsformen der Bedienoberfläche? - Lassen sich funktionale Elemente wie Menüs, Funktionstasten oder Symbolleisten konfigurieren? - Kann der Benutzer Anzeige- und Ausgabeformate den eigenen Bedürfnissen anpassen?
FEHLERTOLERANZ	- Werden Benutzereingaben auf Plausibilität hin geprüft? - Führen Eingabefehler zu Programmabstürzen? - Sind Fehlermeldungen in einer verständlichen Sprache verfasst? - Werden Korrekturhinweise angeboten?

Tabelle 5.1 – Unterkriterien der Grundsätze der ISO 9241-110

5.2 Datenaufbereitung und -integration

Die Aufbereitung von Daten sowie deren anschließende Integration in ein Datawarehouse ist eine elementare Aufgabe von Business Intelligence Systemen. Ziel ist es, Daten aus verschiedensten Quellen mittels Implementierung von ETL-Prozessen in ein einheitliches Format zu bringen und diese zusammenzuführen. Für diesen Zweck stellt Pentaho das

Werkzeug Data Integration zur Verfügung, welches sich als eigenständige Java-Applikation einsetzen lässt.

Für die Aufgabenbewältigung wird dem Benutzer bei der ersten Ausführung eine große, übersichtliche und leere Arbeitsfläche mit einer links angeordneten Baumstruktur präsentiert. Diese ist in verschiedenen Gruppen wie z.B. Input, Output oder Transform unterteilt. Innerhalb dieser befinden sich eine Vielzahl von Schnittstellen zu unterschiedlichen Datenbanken und -formaten (vgl. 7.1), Möglichkeiten der Aufbereitung sowie auch Optionen der Datenbereitstellung wodurch eine hohe Funktionalität gewährleistet wird. Obendrein lässt sich diese durch beliebige Plugins, welche in Form von Java-Archiven eingebunden werden erweitern. Hierfür steht ein interner Plugin-Browser bereit der diverse Plugins, nach Typen gegliedert, zeilenweise auflistet. Durch Anlegen einer neuen Zeile und der entsprechenden Pfadangabe zum Archiv ist ein Plugin so schnell und einfach eingebunden. Um auch große Datenmengen effizient zu verarbeiten werden neben den gängigen Datenbankschnittstellen auch Big Data Lösungen wie Hadoop oder MongoDB angeboten. Des Weiteren steht z.B. auch ein Bulk-Loader für Oracle oder MySQL zur Verfügung. Programmintern werden diese jeweiligen Funktionen dabei als Steps bezeichnet.

Hinsichtlich der Benutzung und die in diesem Zusammenhang zugehörige Lernförderlichkeit wird spätestens bei Auswahl, entweder durch einen Doppelklick oder mittels Drag & Drop, eines Steps die Arbeitsweise dieses Tools durch das Prinzip Learning by Doing deutlich gemacht. Jeder Step erscheint als Symbol auf der Arbeitsfläche und lässt sich nach Belieben mit anderen Steps als Prozessfolge, sogenannte Transformation welche im XML-Format gespeichert werden, verbinden. Durch den immer wiederkehrenden ähnlichen Ablauf der Dialoge und dem Erfüllen des „What You See Is What You Get-Prinzip“ ist dieses Feature auch erwartungskonform. Darüber hinaus beinhaltet jeder Step auch ein Konfigurationsmenü, indem alle relevanten Einstellungen vorgenommen werden können (vgl. 7.2, 7.3 & 7.4).

Bezüglich der Aufgabenangemessenheit wird der Benutzer während der Erstellung weitestgehend vom System unterstützt. Unnötige Arbeitsschritte werden vermieden, indem z.B. bei Eingabefeldern oft sinnvolle Standardwerte vorgegeben werden. So entfällt ein Ausfüllen irrelevanter Formularfelder und es entsteht eine Zeitersparnis. Auch werden größtenteils alle Optionen mittels einem Mouseover und ein dadurch erscheinendes Layer kurz erklärt. Dem Benutzer wird so verständlich gemacht, was von ihm verlangt wird. Wie in Abschnitt 4.1 gezeigt, kann dieser so ohne große Einarbeitungszeit einfache aber auch sehr komplexe Transformationen effektiv und effizient definieren. Für weiterführende Hilfestellungen muss jedoch auf die Webseiten von Pentaho zurückgegriffen werden. Einen direkten Herstellersupport gibt es für die CE nicht. Über den eigenen YouTube Channel, dem Infocenter oder auch der Wiki-Seite lassen sich jedoch zahlreiche Informationen, wie auch Tutorials, finden. Anzumerken ist hier allerdings, dass diese Informationen immer für

die EE gedacht sind. Folglich besteht die Möglichkeit, dass man auf Funktionen trifft, welche in der CE nicht integriert sind. Bei gravierenden Problemen steht auch ein Forum zur Verfügung, in dem Mitglieder der Pentaho Community bereit sind auszuhelfen.

Bezugnehmend auf die Steuerbarkeit können irrtümliche Fehler bei der Eingabe oder ein versehentliches Löschen diverser Steps mit der integrierten Undo-Funktion schnell rückgängig gemacht werden. Die Transformationen können nach Erstellung entweder direkt ausgeführt oder wie in Abschnitt 4.2.3 erwähnt (vgl. auch 7.13) als sogenannte Jobs, zu einer sequentiellen Abfolge verkettet werden. Eine Speicherung dieser sowie die Fortführung der Erstellung zu einem späteren Zeitpunkt ist ebenfalls problemlos möglich. Mit Hilfe einer integrierten Preview Funktion kann darüber hinaus zu jedem Zeitpunkt eine gewünschte Anzahl von Datensätzen angezeigt und so überprüft werden. Während einer Ausführung erhält der Benutzer außerdem in Echtzeit eine Rückmeldung über die durchgeführten Schritte. So werden z.B. auch die nach eigens aufgestellten Kriterien gefilterten Daten beziffert, wodurch sich wie in Abschnitt 4.1 gezeigt eine Statistik erstellen lässt. Überdies werden irreversible Anweisungen erst dann durchgeführt, wenn der Benutzer diese entsprechenden Optionen, wie z.B. das Löschen von Daten in einer Datenbank, vorab manuell aktiviert. Die Selbstbeschreibungsfähigkeit wird daher in vollem Zuge erfüllt.

Darüber hinaus gestaltet sich auch eine mögliche Fehleranalyse als komfortabel. Fehlerhafte Steps werden dabei farblich hervorgehoben und im Protokoll mit der entsprechenden Ursache verständlich vermerkt. So kann z.B. eine fehlerhafte Syntax in einer SQL-Abfrage oder ein Konvertierungsfehler durch einer falschen Datentypangabe schnell korrigiert werden. Anzumerken ist hier jedoch, dass Fehlerhinweise erst nach Ausführung einer Transformationen bzw. eines Jobs angezeigt werden. Auch fehlen konkrete Korrekturvorschläge.

In Bezug auf die Individualisierbarkeit kann der Benutzer über die Programmeinstellungen die Arbeitsfläche nach persönlichen Bedürfnissen gestalten. Mögliche Änderungen sind hier neben der Sprache, die Schriftart sowie –größe, Farbeinstellungen aber auch Default-Werte für interne Programmfunktionen. Lediglich die Symbolleisten bzw. Menüs sind nicht konfigurierbar.

5.3 Datenanalyse

Der Kernteil von Business Intelligence Systemen besteht in der Datenanalyse. Dieser Arbeitsschritt erfolgt bei Pentaho innerhalb der Business Analytics Plattform, welche das Herzstück der Suite darstellt. Hier können zum einen entsprechende Analysen durch verschiedene Werkzeuge erstellt, zum anderen aber auch dessen generierte Ergebnisse für Entscheidungsträger auf dem Server gespeichert und publiziert werden. Die Plattform an sich ist sehr übersichtlich aufgebaut, wodurch der Benutzer sich schnell zurechtfindet.

Neben einer einfachen Benutzerverwaltung mit Vergabe verschiedener Rechte, besteht auch der Zugriff auf einen integrierten Marketplace. Hier werden neben einer Vielzahl durch die Community entwickelter Plugins auch Sprachpakete oder Themes zur Individualisierung der Oberfläche angeboten. Positiv hervorzuheben ist hier auch die Kennzeichnung des Entwicklungsstatus der jeweiligen Plugins durch verschiedene Stages (1-4). Hierdurch wird dem Benutzer ermöglicht potenziell nützliche Plugins zu erproben, welche sich noch im Anfangsstadium der Entwicklung befinden und so ggf. die Funktionalität der Suite zu erweitern. Darüber hinaus ermöglicht das dazugehörige Tool Sparkl, die Entwicklung sowie Publizierung eigener Plugins.

Wie in Abschnitt 4.2 dargestellt, sind für die Datenanalysen die vorab aufbereiteten und integrierten Daten als Quelle anzulegen. Dies erfolgt auf relativ einfachem Wege über den integrierten Data Source Manager indem z.B. eine direkte Datenbankverbindung zu gängigen Datenbanken angelegt wird. Hintergrund für die Analysen ist dabei ein Star-Schema, aus dem ein entsprechender Cube für die Analyse erstellt wird. Eine direkte Verbindung zu einer MongoDB Datenbank wird derzeit jedoch nicht unterstützt. Die Dimensionen der Cubes sowie auch benötigte Kennzahlen lassen sich hierbei über den im Data Source Manager integrierten Data Source Model Editor verwalten (vgl. 7.9). Dieser ist sehr übersichtlich aufgebaut und so für den Benutzer auch leicht zu bedienen. Neben dem Hinzufügen bzw. Entfernen von Dimensionen, lassen sich auch deren Hierarchien verwalten. Die Kennzahlenerstellung ist jedoch auf einfache Aggregationen wie „SUM“, „AVERAGE“, „MINIMUM“, „MAXIMUM“, „COUNT“ oder „COUNT_DISTINCT“ beschränkt. Sollten dem Benutzer diese Aggregationen nicht ausreichen, so kann dieser mittels dem externen Tool Schema Workbench die dem Cube zugrunde liegende XML Datei bearbeiten. Dadurch lassen sich auch solche Kennzahlen erstellen, welche z.B. auf Abfragen mittels SQL zurückgreifen. Zu beachten ist hier, dass bei der Namensgebung erstellter und genutzter Kennzahlen keine Umlaute verwendet werden. Andererseits führt dies beim Wiederöffnen bereits gespeicherter Analysen zu einer NullPointerException.

Für die Aufgabenbewältigung und somit auch die eigentlichen Analysen der Cubes ist in der Basisinstallation das Werkzeug JPivot enthalten. Da aber auch das Saiku Analytics CE Plugin zunehmend an Beliebtheit gewonnen hat, wurde dieses zum Vergleich mit einbezogen. Beide Werkzeuge nutzen als Basis den im Data Source Manager generierten Cube. Wie in den Abschnitten 4.2.3 oder auch 4.2.4 dargestellt, verfügen diese über eine übersichtliche Navigationsleiste, welche alle gängigen Cube-Operationen wie die Pivotierung, den Drill Down, die Dice Funktion oder auch die Split & Merge Funktion beinhaltet. Dem Benutzer wird so ermöglicht die Daten aus unterschiedlichen Perspektiven zu betrachten und diese nach Belieben schnell und einfach zu filtern oder zu sortieren. Dies gewährleistet auch eine gute Fehlertoleranz, da Fehler innerhalb der Anwendungen so weitestgehend nicht mehr entstehen können.

Wohingegen JPivot noch eine manuelle Auswahl der Filter bzw. Kennzahlen erfordert, bietet Saiku Analytics in Bezug auf die Funktionalität bereits ein Drag & Drop und darüber hinaus auch weitergehende, hilfreiche Filtermöglichkeiten, wie z.B. eine begrenzte Ausgabe der Datensätze, an (vgl. 7.16). Bezugnehmend auf die Aufgabenangemessenheit verursacht JPivot durch diese manuelle Auswahl weitere Arbeitsschritte, die insbesondere bei einer hohen Anzahl von Dimensionen vermieden werden könnten. Die interne Bearbeitung der jeweiligen Operationen findet durch ein automatisches Bilden von MDX Abfragen statt. Dabei erhält der Nutzer über ein entsprechendes Symbol in der Navigationsleiste jederzeit Zugriff auf die generierten Abfragen und kann diese bei Bedarf einsehen und zusätzlich anpassen. Allerdings findet in diesem Zusammenhang keine Überprüfung des Codes statt. Die Selbstbeschreibungsfähigkeit ist dabei in vollem Umfang erfüllt. Folgenschwere Handlungen können aufgrund eines reinen Lesezugriffs nicht erfolgen und auch sind darüber hinaus keine weiteren Hilfestellungen für den Benutzer erforderlich. Kennt dieser die Bedeutung der einzelnen Cube Operationen, so sind beide Werkzeuge selbsterklärend und lassen sich ohne Probleme verwenden. Darüber hinaus wird die Lernförderlichkeit durch ein Learning by Doing unterstützt. Ebenfalls existieren Beispielanalysen, denen sich ein Benutzer bedienen kann.

Die Ergebnisausgabe dabei wie in Abschnitt 4.2 beschrieben sowohl bei JPivot als auch bei Saiku Analytics in gängigen Diagrammformen (vgl. 7.10) oder aber auch wie in Abschnitt 4.2.2 gezeigt in Tabellenform erfolgen. Darüber hinaus bietet JPivot im Vergleich zu Saiku Analytics zahlreiche manuelle Anpassungen, wie z.B. die Skalierung, Achsenbeschriftungen oder Farbeinstellungen von Charts. Dennoch fehlen hier neben einigen Varianten der bekannten Diagrammtypen die wichtigen Geo-Charts. Wie in Abschnitt 4.2.1 dargestellt, ließ sich diese Lücke bei Saiku Analytics durch die Installation der Erweiterung Saiku Chart Plus schließen. Mittels dieser wird auf das Kartenmaterial von Google zugegriffen, wodurch es ermöglicht wird ganze Regionen und auch einzelne Punkte auf einer Weltkarte zu visualisieren. Obwohl das gewünschte Ziel aufgrund der Limitierung von maximal 400 Einträgen nicht erreicht werden konnte, ist diese Art von Visualisierung für andere Institutionen, wie z.B. produzierende Unternehmen, eine gute Alternative. So lassen sich z.B. die Verkaufszahlen innerhalb bestimmter Länder wunderbar darstellen (vgl. 7.7). Die erstellten Analysen lassen sich im Anschluss zu einer möglichen weiteren Bearbeitung direkt auf dem Server speichern oder auf Wunsch auch in verschiedene Formate wie z.B. XLS oder PDF exportieren.

Bezüglich der Individualisierbarkeit lassen sich keine direkten Anpassungen der Anwendungen vornehmen. JPivot liegt hier nur in der englischen, Saiku Analytics dagegen in der deutschen Sprache vor. Die Bedienoberfläche kann für die Business Analytics Plattform jedoch angepasst werden. Da diese beiden Werkzeuge auch in dieser integriert sind, wird die Ansicht entsprechend übernommen.

Für das Arbeiten im Big Data Bereich empfiehlt es sich entsprechende Lösungsansätze zu verwenden. Die einfachen Datenbanken wie z.B. MySQL oder Postgre sind zwar auch einsetzbar, können jedoch aufgrund der langsameren Verarbeitung der Daten zu einer Verzögerung führen. Aufgrund der Arbeit in Echtzeit innerhalb der Cubes würde solch eine Verzögerung den Arbeitsablauf stören. Ferner lassen sich in Betracht auf die Steuerbarkeit bereits getätigte Operationen nur innerhalb von Saiku Analytics abbrechen. Sollte ein Bearbeitungsschritt länger als fünf Minuten benötigen, so wird der Vorgang automatisch durch eine TimeOutException unterbrochen. Optional lässt sich diese Einstellung in den Konfigurationsdateien jedoch anpassen. Eine Undo-Funktion ist ebenfalls nicht vorhanden. Diese wird aber auch nicht zwingend benötigt, da sich durch wenige Klicks jede gewünschte Ansicht herstellen lässt. Insgesamt erhält ein Benutzer daher zwei erwartungskonforme Anwendungen für die Datenanalyse.

5.4 Berichtswesen

Die in den Datenanalysen gesammelten Ergebnisse müssen für die Entscheidungsträger in einer schnell verständlichen Form präsentiert werden können. Pentaho bietet dem Benutzer hierfür zwei Möglichkeiten an. Zum einen lassen sich die Ergebnisse in Berichte und zum anderen in Form eines Dashboards präsentieren.

Bezugnehmend auf die Möglichkeit der Berichtserstellung stellt Pentaho für die Aufgabenbewältigung den Report Designer als Anwendung bereit. Die Arbeitsfläche ist hier in verschiedene Bereiche wie z.B. Page Header, Report Header, Group Header oder Details unterteilt. Der Benutzer ist so in der Lage, Objekte per Drag & Drop innerhalb dieser Bereiche pixelgenau zu platzieren. Hierfür wird eine Vielzahl von Möglichkeiten zur Verfügung gestellt. So können z.B., wie in Abschnitt 4.2.2 (vgl. auch 7.11) gezeigt Label, Charts und Ergebnisse von Analysen dargestellt werden. Aber auch andere Objekte wie Texte, Grafiken, Barcodes oder gar ganze Tabellen sind möglich. Alle notwendigen Parameter sind zudem bereits voreingestellt und können vom Nutzer einfach und schnell angepasst werden. Bezüglich der Charts sind alle gängigen Varianten vertreten, welche sich auch individuell sehr detailliert konfigurieren lassen. Lediglich Geo-Charts, welche mittlerweile aber eine wichtige Rolle spielen, sind bislang nicht integriert.

Auch wird hier auf die Funktionalität großen Wert gelegt. Die Objekte können dabei mit fixem Inhalt oder mit parametrisierten Daten, welche sich erst zur Laufzeit füllen, versehen werden. Von letzterer Option wurde wie in den Abschnitten 4.2.1 - 4.2.4 gezeigt auch überwiegend Gebrauch gemacht. Hintergrund ist dabei die Verknüpfung des Berichts mit einer Datenquelle, welche wahlweise z.B. eine einfache Datenbank, eine Tabelle, ein ETL-Prozess aus Pentaho Data Integration oder auch eine Verknüpfung zu einer MongoDB Datenbank sein kann. Für den Fall, dass ein Objekt z.B. mit einer SQL-Abfrage verbunden wird und die Rückgabe mehr als eine Zeile enthält, so ist dieses, wie im Abschnitt 4.2.4 (vgl. auch 7.17) gezeigt, zwingend im Group Header oder in den Details zu platzieren. In allen

anderen Bereichen ist die Ausgabe auf eine Zeile limitiert und daher eher für fixe Inhalte geeignet. Da diese Information nicht direkt ersichtlich ist und extern bezogen werden musste, schränkt dies die Selbstbeschreibungsfähigkeit und zum Teil auch die Erwartungskonformität der Anwendung leicht ein. Auch ist die Hilfestellung seitens Pentaho nur eingeschränkt verfügbar. Zudem ist eine Erweiterung der Funktionalität mittels Plugins derzeit nicht möglich.

Das eigentliche Bedienprinzip des Report Designers wird dem Benutzer erst nach einiger Einarbeitungszeit klar gemacht. Da die Anwendung aber durchaus Ähnlichkeit mit einem Text- oder Bildbearbeitungsprogramm hat und sich die Arbeitsschritte wiederholen, hält sich dieser Aufwand in Grenzen. Darüber hinaus sind in der Anwendung auch Beispielreports zu finden, welche dem Benutzer einige Funktionen vorführen und so auch die Lernförderlichkeit weiter unterstützen. Optional lässt sich auch der Report Design Wizard, welcher einen integrierten Assistenten darstellt, nutzen. Mittels diesem wird der Benutzer Schritt für Schritt bei der Erstellung eines Berichts begleitet, wodurch auch technisch weniger versierte Benutzer in der Lage sind schnell und einfach Berichte zu erstellen. Auch besteht hier die Wahl eines bereits vorbereiteten Designs. Nach Erstellung kann dieses, sofern nötig, ebenfalls manuell weiter bearbeitet werden.

In Bezug auf die Aufgabenangemessenheit ist hier anzumerken, dass ein Bericht immer nur mit genau einer Datenquelle verbunden werden kann. Wie in Abschnitt 4.2.1 (vgl. auch 7.8) dargestellt, ist eine Datenquelle oft nicht ausreichend. Umgehen lässt sich dies nur mittels Nutzung von sogenannten Subreports. Diese stellen im Grunde genommen jeweils einen eigenen, neuen Bericht, jedoch mit einer anderen Datenquelle, dar und können nach Belieben im eigentlichen Report angeordnet werden. Sind eine Vielzahl unterschiedlicher Datenquellen notwendig, so kann sich die Berichtserstellung etwas mühsam gestalten. Dennoch kann der Benutzer alle Aspekte der Anwendung frei steuern. So ist auch eine Zwischenspeicherung der Ergebnisse für eine spätere Weiterbearbeitung jederzeit möglich. Positiv anzumerken ist ebenfalls, dass es dem Benutzer zu jeder Zeit gestattet wird eine Vorschau der bisherigen Berichtsentwicklung anzuzeigen. Abgeschlossene Berichte können wahlweise als HTML, PDF, RTF, CSV oder Excel-Datei gespeichert und manuell verteilt oder direkt auf dem BI-Server publiziert werden.

Überdies verfügt das Feature über eine gute Fehlertoleranz. So wird auf etwaige Fehlermeldungen, wie z.B. eine fehlende Verbindung zur Datenbank, sowie wichtige Hinweise dezent über zwei Symbole hingewiesen ohne das die Arbeitsschritte unterbrochen werden müssen. Nach Aufruf erhält der Benutzer über ein Untermenü alle Meldungen nachvollziehbar aufgelistet und kann sich diesen aneignen. Konkrete Korrekturvorschläge sind jedoch bislang nicht integriert.

Des Weiteren wird auch ein gewisser Grad der Individualisierung eingeräumt. Dies beinhaltet sowohl alle möglichen Farb-, Schrift- und Objekteinstellungen, aber auch andere

wie z.B. ein persönliches Look & Feel. Die Sprache der Oberfläche ist jedoch nur auf Englisch verfügbar und lässt sich nicht anpassen. Auch sind die funktionalen Elemente, wie Menüs oder Symbolleisten fixiert und nicht konfigurierbar.

Die Erstellung eines Dashboards dagegen kann in der CE nur mittels dem Community Dashboard Editor erfolgen, welcher über den BI-Server ausgeführt wird. Obwohl dieser keine Eigenentwicklung von Pentaho ist, ist die Anwendung ein fester Bestandteil der Suite. Anstelle von starren Berichten ist es so möglich für die Entscheidungsträger einen interaktiven Bericht zu erstellen, der auch einen tieferen Informationsgehalt enthält.

Wie in Abschnitt 4.2.5 gezeigt muss der Benutzer für die Aufgabenbewältigung ein Layout (vgl. 7.18), welches aus Zeilen und Spalten besteht, entwerfen. Innerhalb diesem können dann alle denkbaren Komponenten, wie z.B. Charts, Buttons, diverse Menüführungen oder auch Texte integriert werden (vgl. 7.20). Diese können dabei zu jeder Zeit individuell über eine Vielzahl von Einstellungsmöglichkeiten bearbeitet werden. Leider kann der Benutzer bei der Layouterstellung schnell den Überblick verlieren. Dies ist insbesondere dann der Fall, wenn eine Vielzahl von Zeilen und Spalten verwendet wird. Die integrierte Vorschaufunktion hilft nur bedingt weiter, da nur Zeilen und Spalten mit Inhalt klar erkennbar sind. Eine kleine Abhilfe schafft hier die Option des Vergebens von eindeutigen Bezeichnungen der jeweiligen Elemente.

Hinsichtlich der Funktionalität sind die Komponenten wiederum mit einer Datenquelle zu verbinden, welche z.B. das Ergebnis einer SQL bzw. MDX Abfrage oder ein ETL-Prozess sein kann. Auch hier stehen ebenfalls keine Geo-Charts zur Verfügung. Des Weiteren steht derzeit für die Big Data Lösung MongoDB keine direkte Datenbankanbindung zur Verfügung. Sollte mit diesem Konzept gearbeitet werden, ist jede Datenquelle als einen eigenständigen ETL-Prozess in Pentaho Data Integration zu erstellen, was für den Ersteller einen Mehraufwand verursacht. Folglich ist die Aufgabenangemessenheit nicht in vollem Umfang erfüllt. Zwar hat der Benutzer in der Gestaltung des Dashboards freie Wahl, doch muss dann vieles, wie z.B. ein individuelles Design (vgl. 7.19), manuell implementiert werden. So bleibt ein Dashboard mit hohen Ansprüchen nur Entwicklern vorbehalten, da Kenntnisse in HTML, CSS, SQL, MDX oder Java nötig sind. Dementsprechend ist diese Anwendung auch nicht erwartungskonform.

Ansonsten ist die Arbeitsfläche des CDE übersichtlich in drei Bereiche – Layout, Komponenten und Datenquellen – unterteilt, welche wiederrum die enthaltenen Features in Gruppen übersichtlich gliedert. Der Benutzer kann sich so gut zurechtfinden und erfährt schnell, wo eine Eingabe erwartet wird. Dennoch wird die Selbstbeschreibungsfähigkeit nur zum Teil erfüllt, da der Benutzer ohne vorherige Kenntnisse der Arbeitsweise zu Beginn mit der Vielzahl an gelisteten Elementen und damit verbundenen Funktionen überfordert wird. Folglich ist hier ein etwas höherer Einarbeitungsaufwand von Nöten. Auch stehen hier keine großartigen Hilfeseiten seitens des Herstellers zur Verfügung. Lediglich ein Quickstart

beschreibt kurz und knapp die grundlegenden Funktionen des CDE. Darüber hinaus kann der Benutzer sich einige Demos online anschauen, welche aber nicht die eigentlichen Arbeitsschritte erklären. Hilfreiche Tutorials sind derzeit leider nur kostenpflichtig erhältlich, wodurch die Lernförderlichkeit erheblich eingeschränkt wird.

Obwohl im Hinblick auf die Steuerbarkeit sich Arbeitsschritte unterbrechen und zu einem anderen Zeitpunkt fortführen lassen, müssen irreversible Anweisungen nicht bestätigt werden. Da auch keine Undo-Funktion integriert ist, können versehentliche Durchführungen solcher Aktionen fatal sein. Außerdem erfolgt auch keine Berücksichtigung der Individualisierbarkeit. Die Anwendung steht nur auf Englisch zur Verfügung und auch ein Zuschneiden auf persönliche Bedürfnisse wird hier nicht gestattet.

Zusätzlich wird die Fehlertoleranz hier nahezu ignoriert. Mögliche Fehler muss der Benutzer oft selbst herausfinden, da dieser z.B. bei falsch gesetzten Einstellungen oder etwa falsch verbundenen Komponenten nicht darauf aufmerksam gemacht werden kann. Lediglich die Fehlerquelle lässt sich in diesem Fall aufgrund der strikten Trennung der Komponenten einschränken. Ein Weiterarbeiten ist hier trotz etwaiger Fehler dennoch möglich.

5.5 Fazit

Pentaho bietet Nutzern in Bezug auf Business Intelligence eine mächtige Suite an. Das Open Source Paket deckt dabei mit den auch als allein nutzbaren Modulen für die Datenaufbereitung sowie –integration, die Datenanalyse und das Reporting alle Funktionsbereiche ab.

Insbesondere Pentaho Data Integration glänzt durch die Vielzahl an verschiedenen Möglichkeiten des Datenzugriffs sowie dessen Aufbereitung und dem relativ geringen Interaktionsaufwand. Dadurch können die unterschiedlichsten Anforderungen der Benutzer berücksichtigt werden. Folglich sind entsprechende Analysen nicht nur, wie in dieser Ausarbeitung gezeigt, auf das Transportwesen beschränkt sondern können sich auf alle denkbaren Gebiete ausweiten. Zu nennen sind hier z.B. mögliche Produktionsoptimierungen in der Automobilindustrie durch Auswertung von Maschinen- bzw. Sensordaten, eine verstärkte Unterstützung der inneren Sicherheit durch die Analysen sozialer Netzwerke, die Entstehung neuer Produktideen durch Trendanalysen oder einfach eine gezielte Kostenreduzierung in bestimmten Sektoren durch Analyse dieser zugrunde liegenden Daten.

Nichts desto trotz sind hier Defizite in Bezug auf die Benutzung und darauf bezogen die Lernförderlichkeit, die Individualisierbarkeit sowie die Fehlertoleranz erkennbar. So sind bei komplexen Bearbeitungsvorgängen und eine etwaige benötigte Hilfestellung nur eingeschränkte Mittel möglich. Darüber hinaus sind die funktionalen Elemente wie Menüs

oder Symbolleisten nicht anpassbar. Ferner fehlen bei Fehlermeldungen konkrete Korrekturvorschläge. Auch ist eine Prüfung der Benutzereingaben auf Plausibilität nicht vorhanden, welche hier durchaus hilfreich sein kann.

Die Datenanalyse erfolgt dabei innerhalb der Suite fast ausschließlich auf Basis relationaler Datenbanken. Auch die Verarbeitung großer Datenmengen ist durch den Support der Big Data Konzepte wie z.B. Hadoop gewährleistet. Lediglich an einem vollständigen Support von MongoDB hapert es noch. Ein großer Vorteil ist hier die einfache Erweiterbarkeit neuer Funktionen. Den Anwendern wird hier mittels einem integrierten Marketplace die Möglichkeit geboten, die von der Community entwickelten Plugins kostenlos zu beziehen oder gar eigene zu entwickeln und diese zu publizieren. Trotz dieses Vorzugs ist der Anwender aufgrund der eingeschränkten Features in der CE jedoch schon fast gezwungen auf Plugins auszuweichen. Zu erwähnen ist in diesem Zusammenhang auch, dass für die CE keine Updates angeboten werden. Sofern Bugs vorhanden sind, ist der Anwender hier gezwungen die Fehler entweder eigenständig zu beseitigen oder auf ein neues Release zu warten.

Überdies lässt sich mit Hilfe des Pentaho Report Designers ein statisches Berichtswesen etablieren, wodurch eine Vielzahl von Personen mit Informationen versorgt werden können. Alternativ besteht Dank des integrierten Community Dashboard Editors von webdetails auch die Möglichkeit einer interaktiven Berichtserstattung.

Zusammenfassend ist das Pentaho Open Source Business Intelligence System durchaus ein erfolgsversprechender Ansatz ein Unternehmen in Bezug auf operativer sowie strategischer Entscheidungen zu unterstützen und so dessen Profitabilität zu steigern. Dennoch muss man sich bei Nutzung dieses Pakets im Klaren sein, dass seitens Pentaho keinerlei Support angeboten wird. Bei etwaigen Fragen oder Problemen ist man so auf sich allein gestellt und kann nur mit Hilfe der für die EE angebotenen Supportseiten versuchen eine geeignete Lösung zu finden. Alternativ kann jedoch auch die Community über Foren befragt werden. Dementsprechend entstehen trotz der kostenfreien Software neben möglicher Hardwareanschaffungen weitere betriebliche Kosten durch Nutzung interner Ressourcen für die Einrichtung, Wartung, Prozessmodellierung und anderweitigen Arbeiten innerhalb der Suite. Die nachfolgende Abbildung 5.2 fasst die verwendeten Features in einer Nutzwertanalyse zusammen. Dabei werden die in Abschnitt 5.1 aufgestellten Kriterien berücksichtigt und deren Erfüllung, wie in den Abschnitten 5.2-5.4 dargestellt, bewertet. Jedes beschriebene Defizit mindert dabei entsprechend die zu erreichende Punktzahl. Besonders auffällig ist, dass der Grad der Individualisierung oft schlecht abschneidet. Dies ist aber der Tatsache geschuldet, dass BI-Software für bestimmte Aufgabenstellungen entwickelt wird und auf eine persönliche Individualisierbarkeit selten Rücksicht genommen wird.

Gewichtung Untergewichte (10 = 100%)			Datenaufbereitung & -integration			Datenanalyse						Berichtswesen					
			Pentaho Data Integration			Jpivot			Saiku Analytics			Pentaho Reporting			CDE		
			Erreichte Punkte		Wert	Erreichte Punkte		Wert	Erreichte Punkte		Wert	Erreichte Punkte		Wert	Erreichte Punkte		Wert
Wichtigkeitsgrad			-3-		-4-	-5-		-6-	-7-		-8-	-9-		-10-	-11-		-12-
Feld	-1-	-2-	30%		Feld2 x Feld3	20%		Feld2 x Feld5	20%		Feld2 x Feld7	15%		Feld2 x Feld9	15%		Feld2 x Feld11
Berechnungsformel																	
1. Aufgabenbewältigung	2,9	3	10		30	8		24	9		27	8		24	7		21
2. Funktionalität	2,9	3	10		30	7		21	8		24	8		24	7		21
3. Benutzung	4,2																
3.1 Aufgabenangemessenheit		0,6	10		6	8		4,8	9		5,4	7		4,2	6		3,6
3.2 Selbstbeschreibungsfähigkeit		0,6	10		6	10		6	10		6	8		4,8	6		3,6
3.3 Lernförderlichkeit		0,6	8		4,8	10		6	10		6	6		3,6	4		2,4
3.4 Steuerbarkeit		0,6	10		6	7		4,2	8		4,8	10		6	6		3,6
3.5 Erwartungskonformität		0,6	10		6	9		5,4	10		6	7		4,2	6		3,6
3.6 Individualisierbarkeit		0,6	8		4,8	2		1,2	2		1,2	7		4,2	0		0
3.7 Fehlertoleranz		0,6	7		4,2	8		4,8	8		4,8	8		4,8	2		1,2
Gesamtpunkte (Summe Wert x Wichtigkeitsgrad)					**29,34**			**15,48**			**17,04**			**11,97**			**9**
Gesamtpunkte BI Suite:					**82,83**												

Abbildung 5.2 – Nutzwertanalyse des Pentaho OS BI Systems

6. Zusammenfassung

Innerhalb dieser Arbeit wird gezeigt, dass das Open Source System von Pentaho für Unternehmen aller Art eine potenzielle Lösung darstellt einen großen Datenbestand aufzubereiten, auszuwerten und in einer leicht verständlichen Form zu präsentieren.

Neben der eingehenden Behandlung der zu Grunde liegenden Thematik von Business Intelligence Systemen und den damit in Verbindung stehenden Begriffen wie ETL-Prozessen, OLAP oder Data Mining werden auch die Begriffe Big Data sowie Open Source näher erläutert.

Im weiteren Verlauf werden die wesentlichen Grundelemente des Pentaho Open Source Business Intelligence System erklärt. Da diese modular aufgebaut sind, können sie optional auch eigenständig verwendet werden. Darüber hinaus wird auch die Mitbewerbersituation geschildert. Hier sind lediglich der deutsche Anbieter Jedox sowie der amerikanische Anbieter Jaspersoft zu nennen, welche wie Pentaho auch eine komplette BI Suite anbieten.

Für die eigentliche Evaluation finden praxisnahe Big Data-Anwendungsszenarien Verwendung, welche die Unternehmensbereiche Marketing, Service, Logistik, Personal und Controlling berücksichtigt. Dabei wird auch der Datenursprung erläutert sowie eine Datenprüfung anhand der mit dem Begriff Big Data dargestellten Qualitätsmerkmalen durchgeführt.

Die anschließende Konzeption sowie Realisierung der jeweiligen Szenarien findet innerhalb der Pentaho Anwendung statt. Für diesen Zweck werden die Daten vorab aufbereitet und in eine MySQL Datenbank integriert. Abschließend werden die verwendeten Features anhand der Kriterien Aufgabenbewältigung, Funktionalität sowie Benutzung hinreichend untersucht und bewertet. Diese Bewertung wird außerdem in einer Nutzwertanalyse dargestellt, in welcher für die BI Suite eine Gesamtwertung berechnet wird.

6.1 Ausblick

Betrachtet man das in Abschnitt 2.2 dargestellte exponentielle Wachstum von Big Data, so lässt sich daraus schließen, dass zukünftige operative sowie strategische Entscheidungen in Unternehmen immer schwieriger zu treffen sind. Hintergrund ist der damit verbundene Aufwand aus einer immer größer werdenden Datenmenge in kurzer Zeit eine

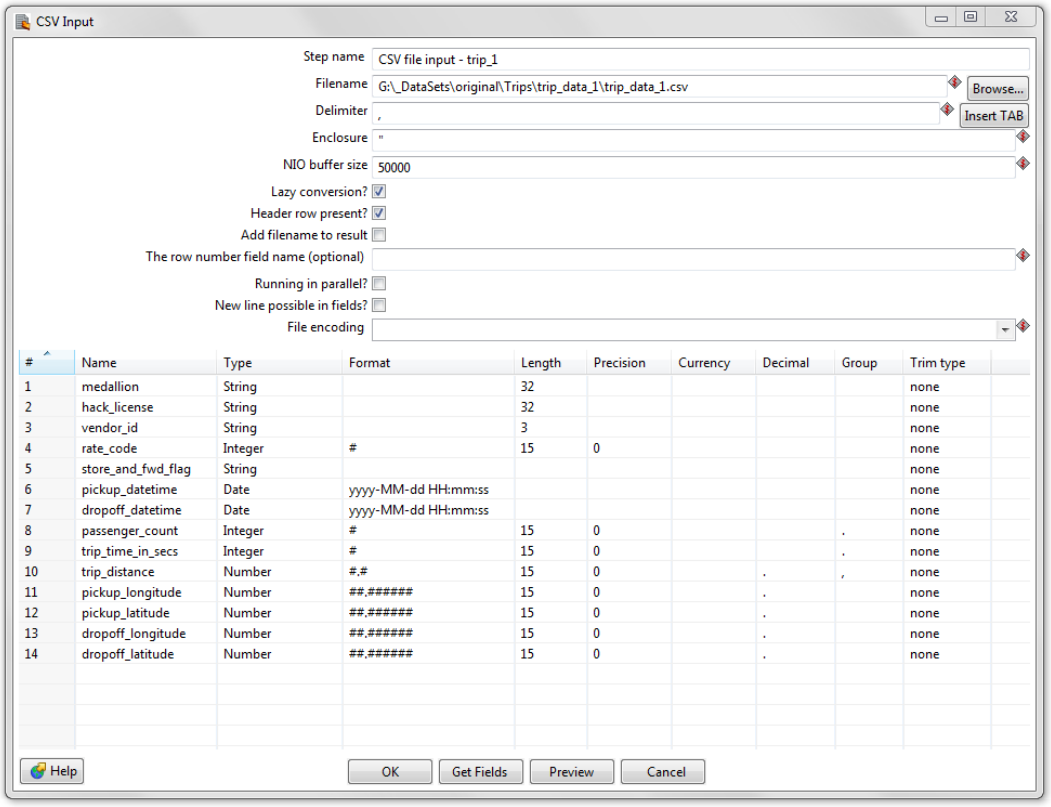
aussagekräftige Information zu bilden. Zugleich bergen diese Daten aber für viele Branchen ein großes Potenzial.

Gemäß dem Bericht „Big data: The next frontier for innovation, competition, and productivity" des McKinsey Global Institute schafft Big Data mehr Transparenz, wodurch Unternehmen einen besseren Überblick erhalten und es ihnen so ermöglicht wird flexibler Entscheidungen zu treffen. Darüber hinaus gestattet solch eine große Datenmenge mehr Planspiele womit auch neue Geschäftsmodelle, Produkte und Dienstleistungen entstehen können. Obendrein können diese aufgrund des verbesserten Zugangs zum einzelnen Kunden auch besser zugeschnitten werden. Dadurch kann die Wertschöpfung optimiert werden (Manyika, et al. 2011).

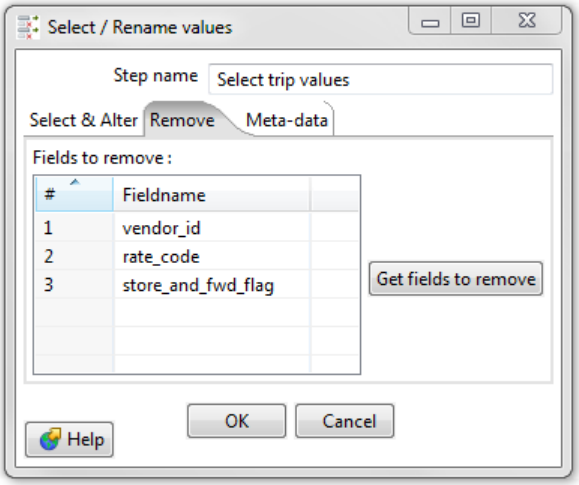
Die logische Konsequenz daraus ist, dass Unternehmen in naher Zukunft ohne BI Software, welche in der Lage sind solche Datenbestände auszuwerten und entsprechend zu präsentieren, nicht mehr auskommen werden. Folglich wird sich der Markt neben den bereits vorhandenen kommerziellen Lösungen sowie Open Source Alternativen aber auch die technischen Lösungen weiter entwickeln und sich so den Anforderungen der Unternehmen anpassen.

7. Anhang

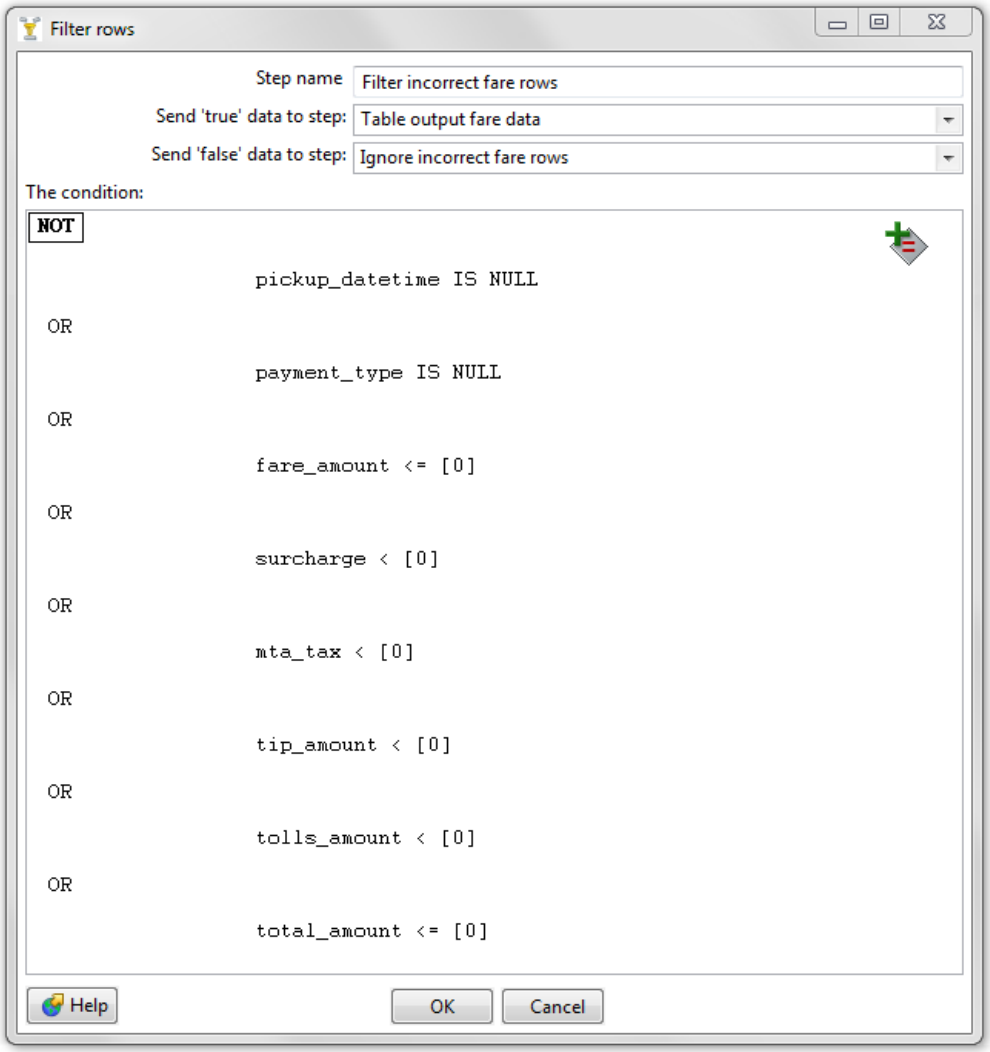
7.1 CSV Input der Reisedaten für den Monat Januar



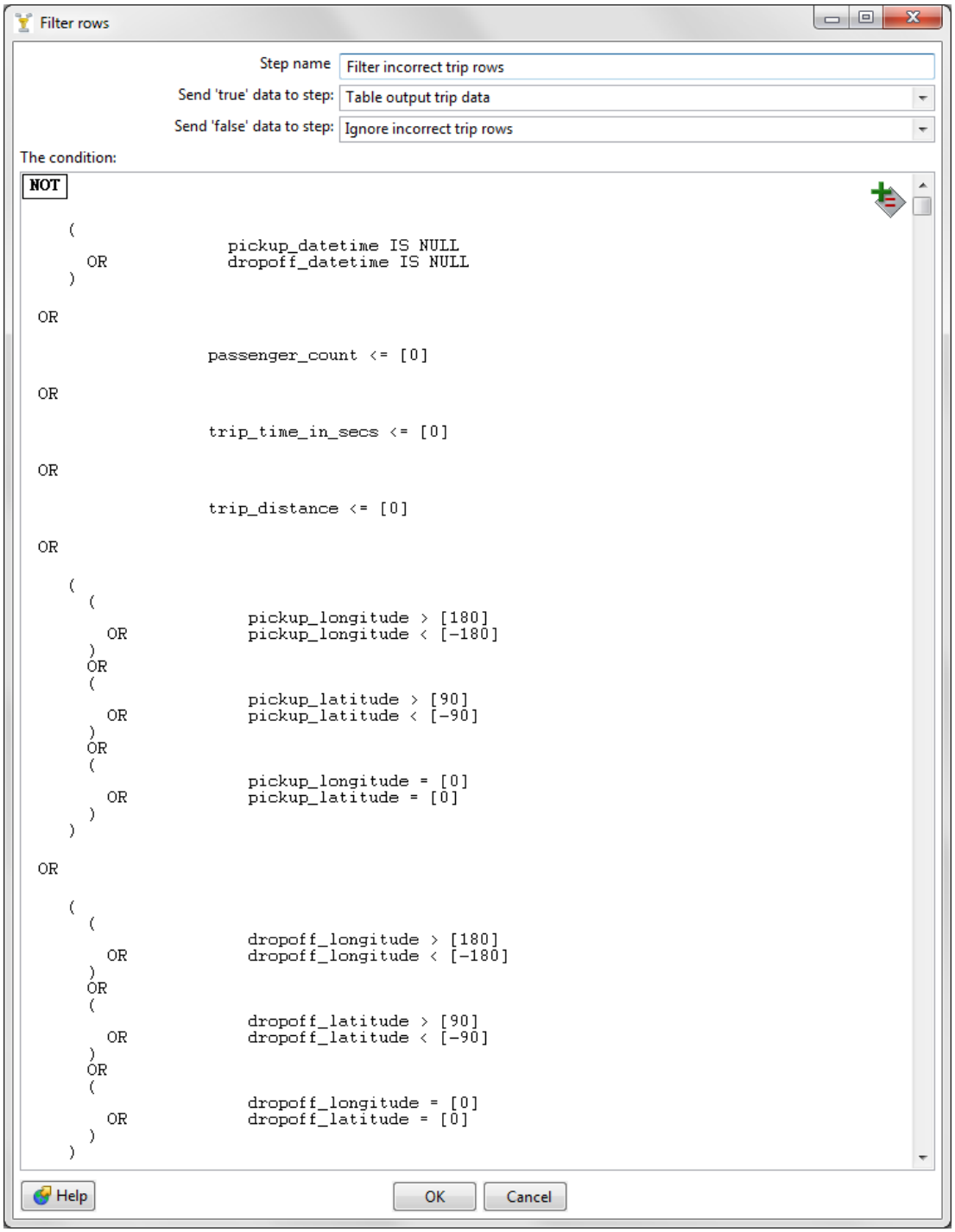
7.2 Select / Rename Values Option für die Reisedaten



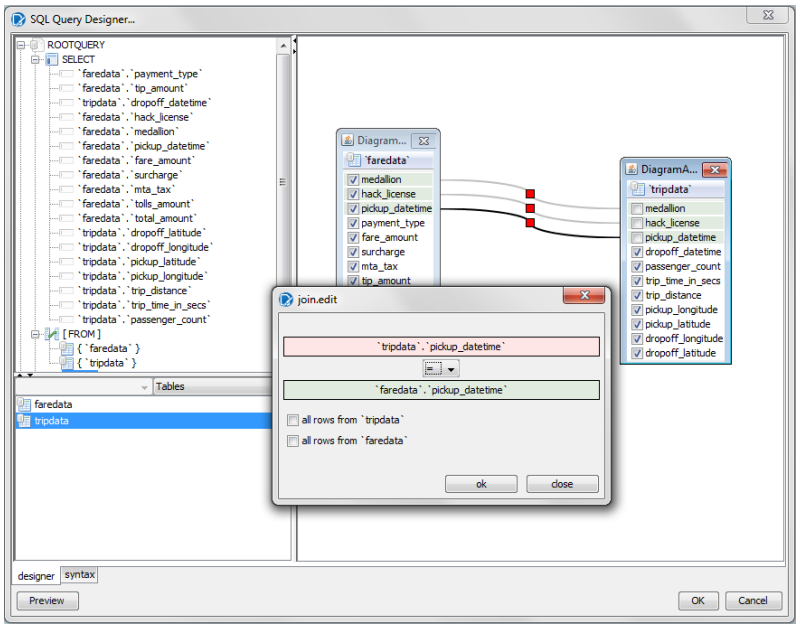
7.3 Filterung fehlerhafter Tarifdaten



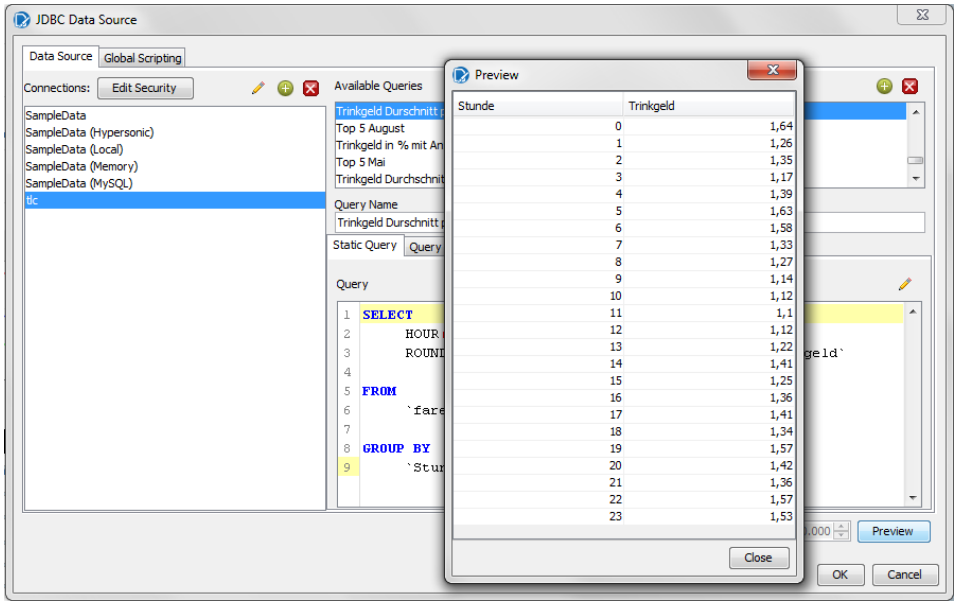
7.4 Filterung fehlerhafter Reisedaten



7.5 SQL Query Designer der Anwendung Pentaho Report Designer

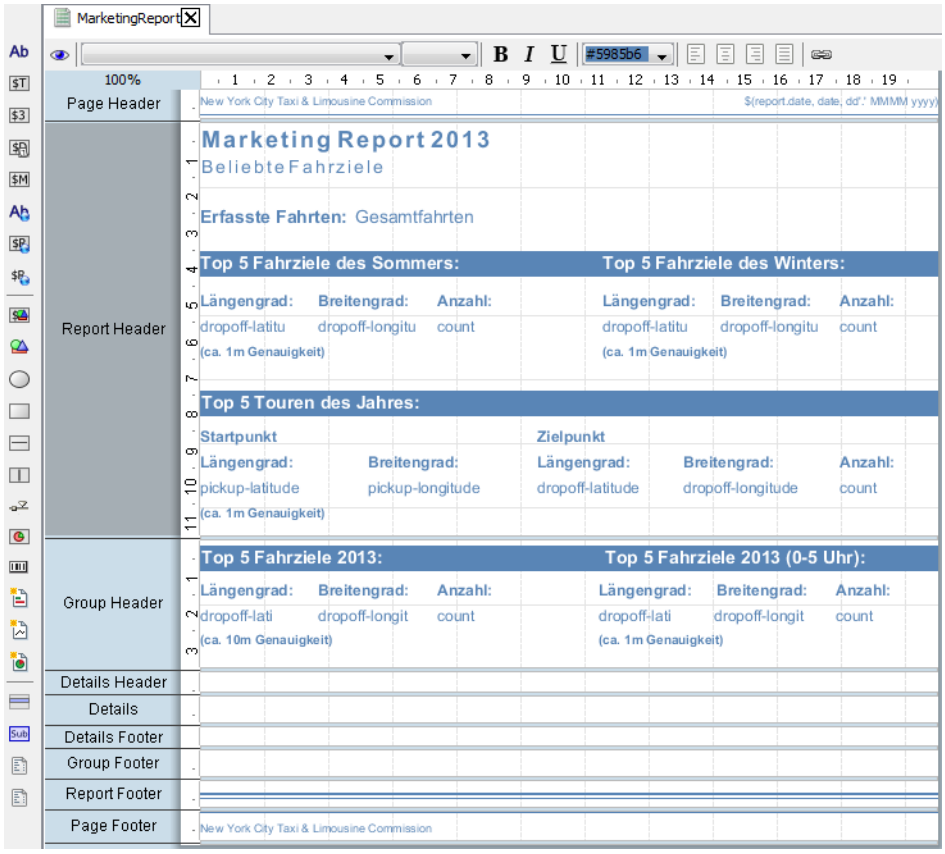


7.6 Ergebnisvorschau einer erstellten SQL-Abfrage

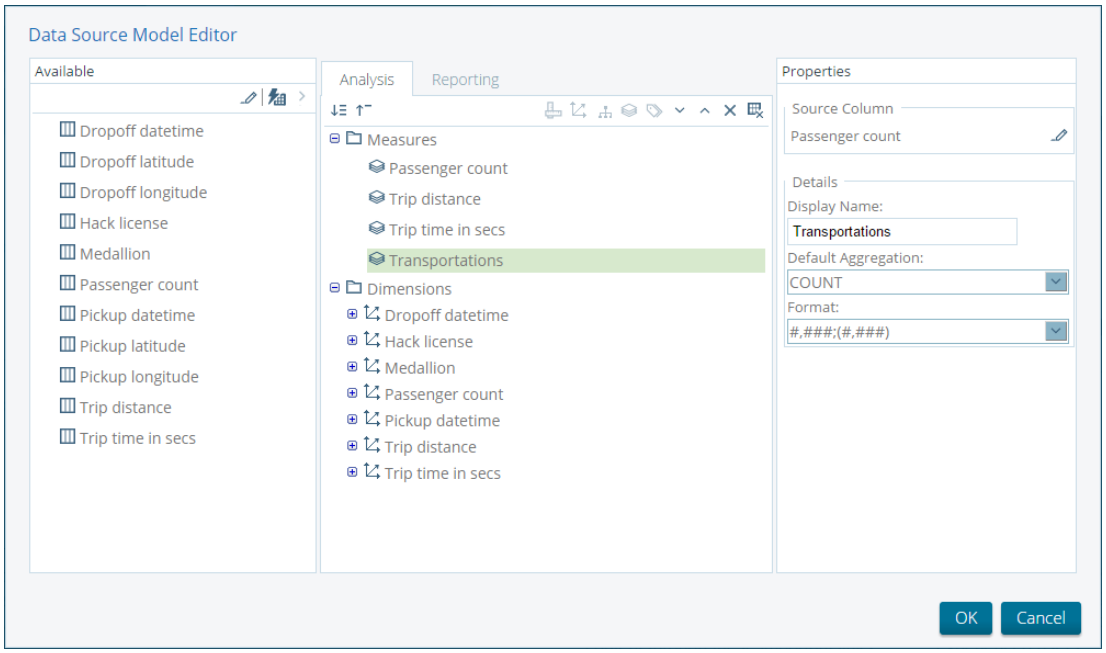


7.7 Visualisierung von Verkaufsdaten mit Saiku Charts Plus

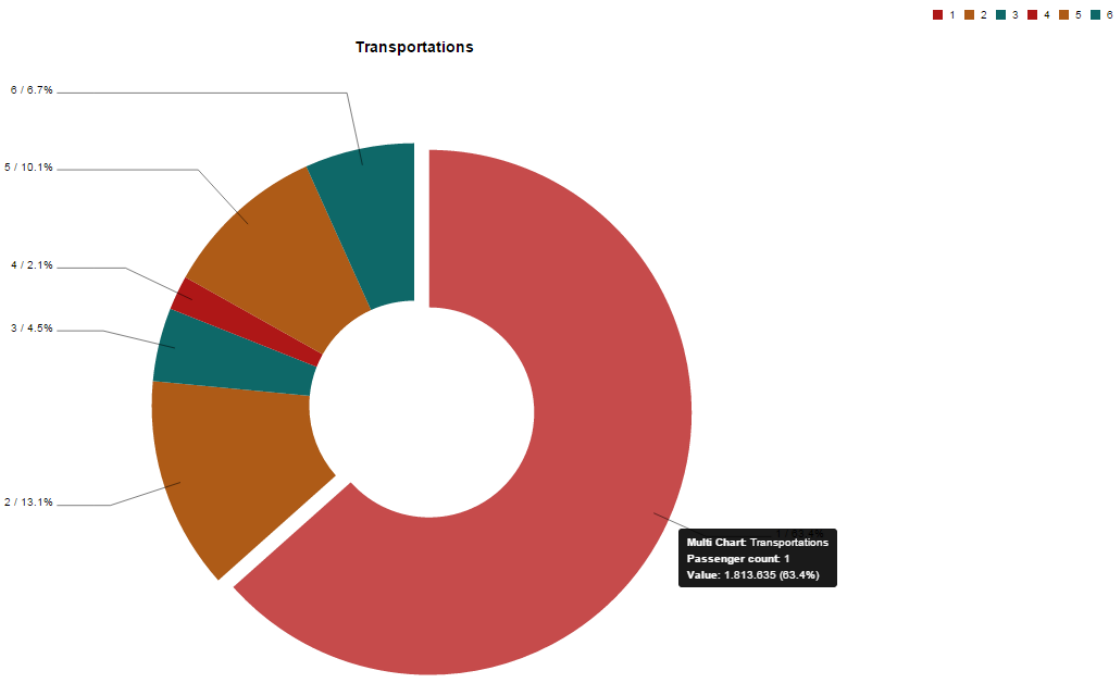
7.8 Entwurfsmodus der Marketinganalyse



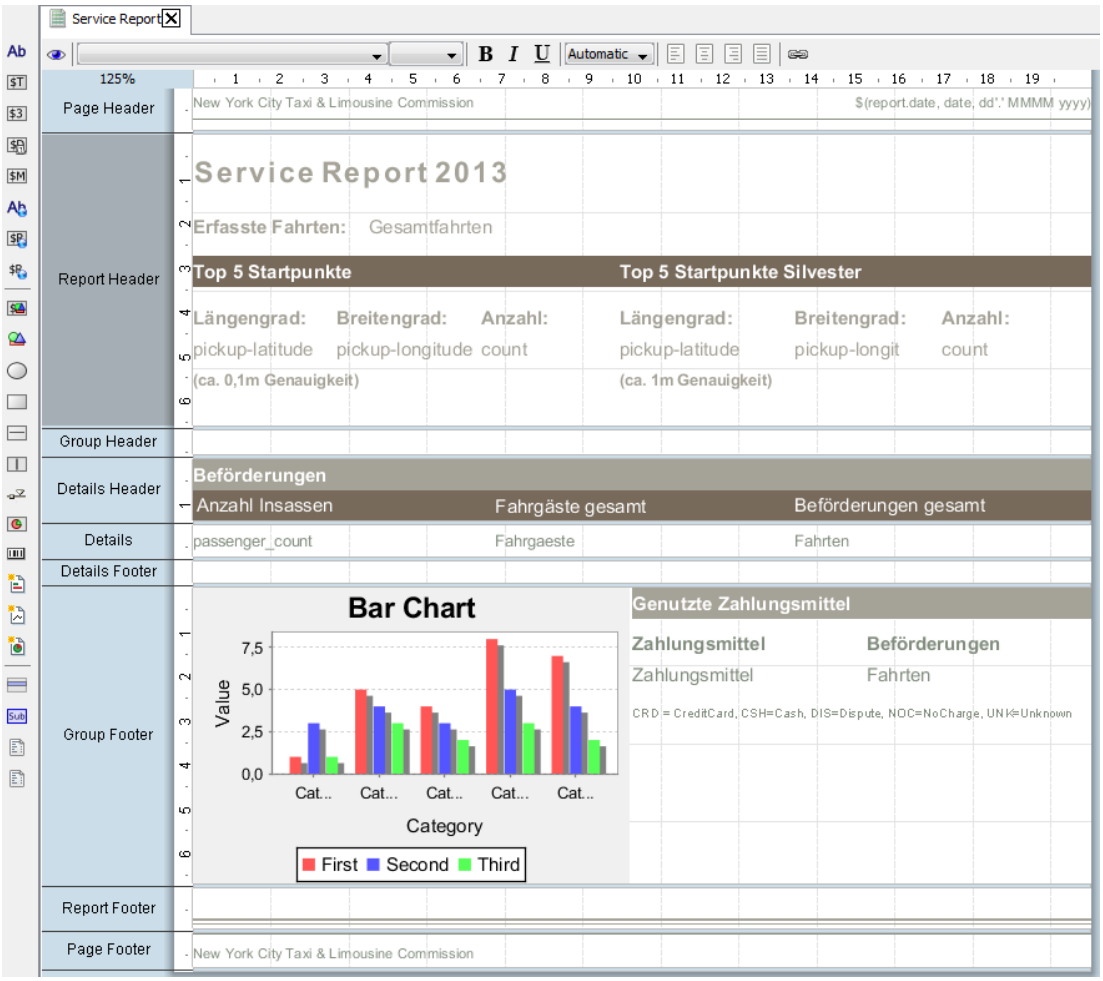
7.9 Erstellung von Kennzahlen in der Business Analytics Plattform



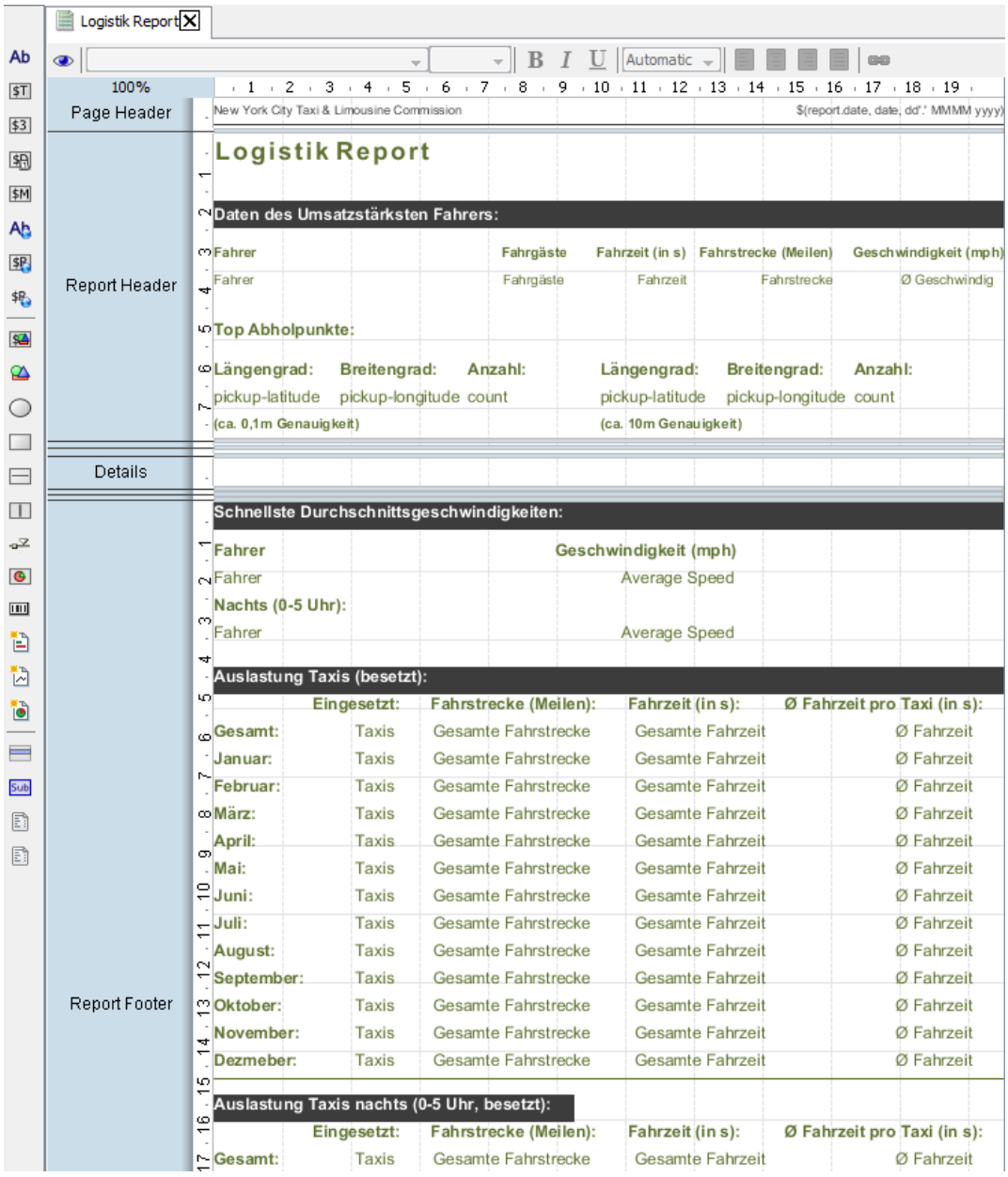
7.10 Tortendiagramm – Analyse der Fahrzeugkapazitäten



7.11 Entwurfsmodus der Serviceanalyse



7.12 Entwurfsmodus der Logistikanalyse



7.13 ETL-Prozess für die Geschwindigkeiten der Fahrer

Execution results

History | Logging | Job metrics | Metrics

Job / Job Entry	Comment	Result	Reason	Filename	Nr	Log date
speed						
Job: speed	Start of job execution		start			2015/03/20 10:21:46
START	Start of job execution		start			2015/03/20 10:21:46
START	Job execution finished	Success			0	2015/03/20 10:21:46
Add speed data	Start of job execution		Followed unconditional link	F:\DataSets\transformations\aktu...		2015/03/20 10:21:46
Add speed data	Job execution finished	Success		F:\DataSets\transformations\aktu...	1	2015/03/20 10:24:34
Control speed	Start of job execution		Followed link after success	file:///F:/DataSets/transformation...		2015/03/20 10:24:34
Control speed	Job execution finished	Success		file:///F:/DataSets/transformation...	2	2015/03/20 10:24:50
Job done	Start of job execution		Followed link after success			2015/03/20 10:24:50
Job done	Job execution finished	Success			2	2015/03/20 10:24:50
Job: speed	Job execution finished	Success	finished		2	2015/03/20 10:24:50

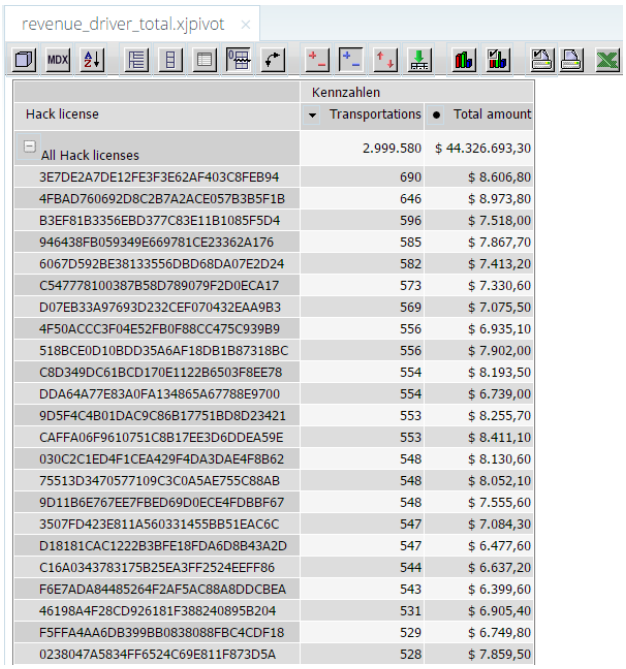
7.14 Ergebnisdarstellung der Logistikanalyse (Seite 2/2)

April:	7.423	712.493,70	171.009.170,00	701,46
Mai:	6.518	793.563,30	180.472.920,00	749,00
Juni:	6.927	706.957,60	179.953.251,00	747,20
Juli:	7.117	716.596,80	186.120.263,00	795,15
August:	4.751	670.966,30	166.676.507,00	702,08
September:	9.203	736.001,50	195.070.047,00	828,12
Oktober:	6.265	777.808,90	196.806.070,00	832,19
November:	8.621	734.974,00	178.324.391,00	762,06
Dezmeber:	6.914	693.105,50	184.724.354,00	808,00

Auslastung Taxis nachts (0-5 Uhr, besetzt):

	Eingesetzt:	Fahrstrecke (Meilen):	Fahrzeit (in s):	Ø Fahrzeit pro Taxi (in s):
Gesamt:	12.183	1.242.666,80	232.371.977,00	679,02
Januar:	4.679	74.780,40	12.152.441,00	701,72
Februar:	5.710	178.012,70	37.213.759,00	679,37
März:	1.386	114.025,40	21.020.635,00	651,76
April:	5.935	170.947,40	34.966.890,00	691,40
Mai:	3.833	66.383,80	10.587.300,00	664,70
Juni:	4.057	72.060,00	13.473.343,00	712,08
Juli:	4.057	72.060,00	13.473.343,00	712,08
August:	3.402	100.136,20	19.094.848,00	635,84
September:	4.120	66.815,10	10.804.297,00	677,56
Oktober:	5.002	121.677,00	24.187.299,00	729,35
November:	6.079	154.535,30	27.593.007,00	664,75
Dezmeber:	2.302	33.906,40	5.838.499,00	645,42

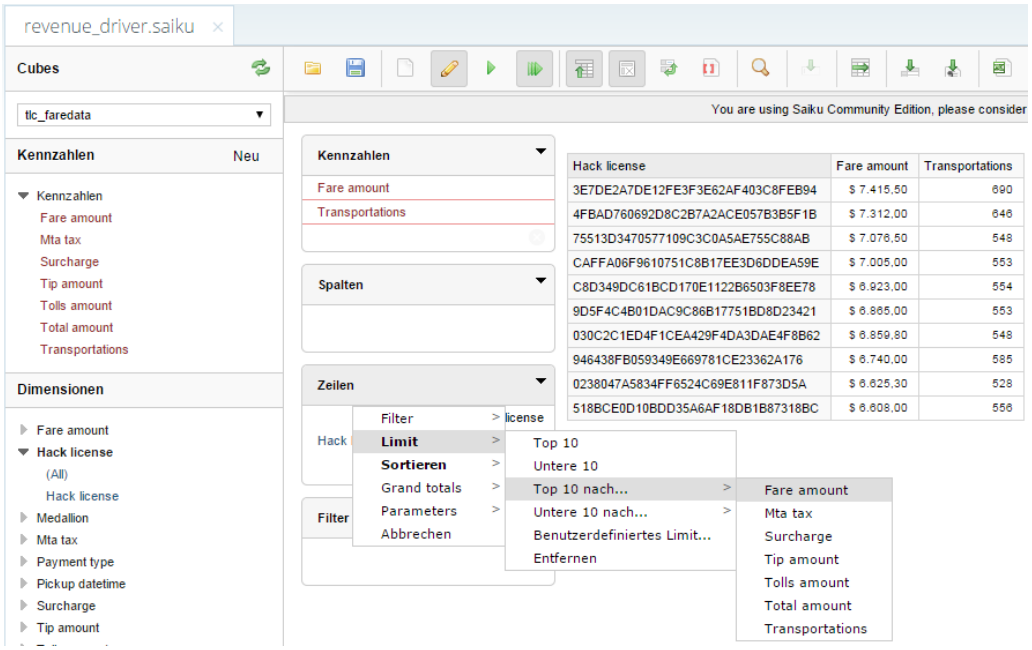
7.15 Analyse der umsatzstärksten Fahrer anhand der Gesamteinnahmen



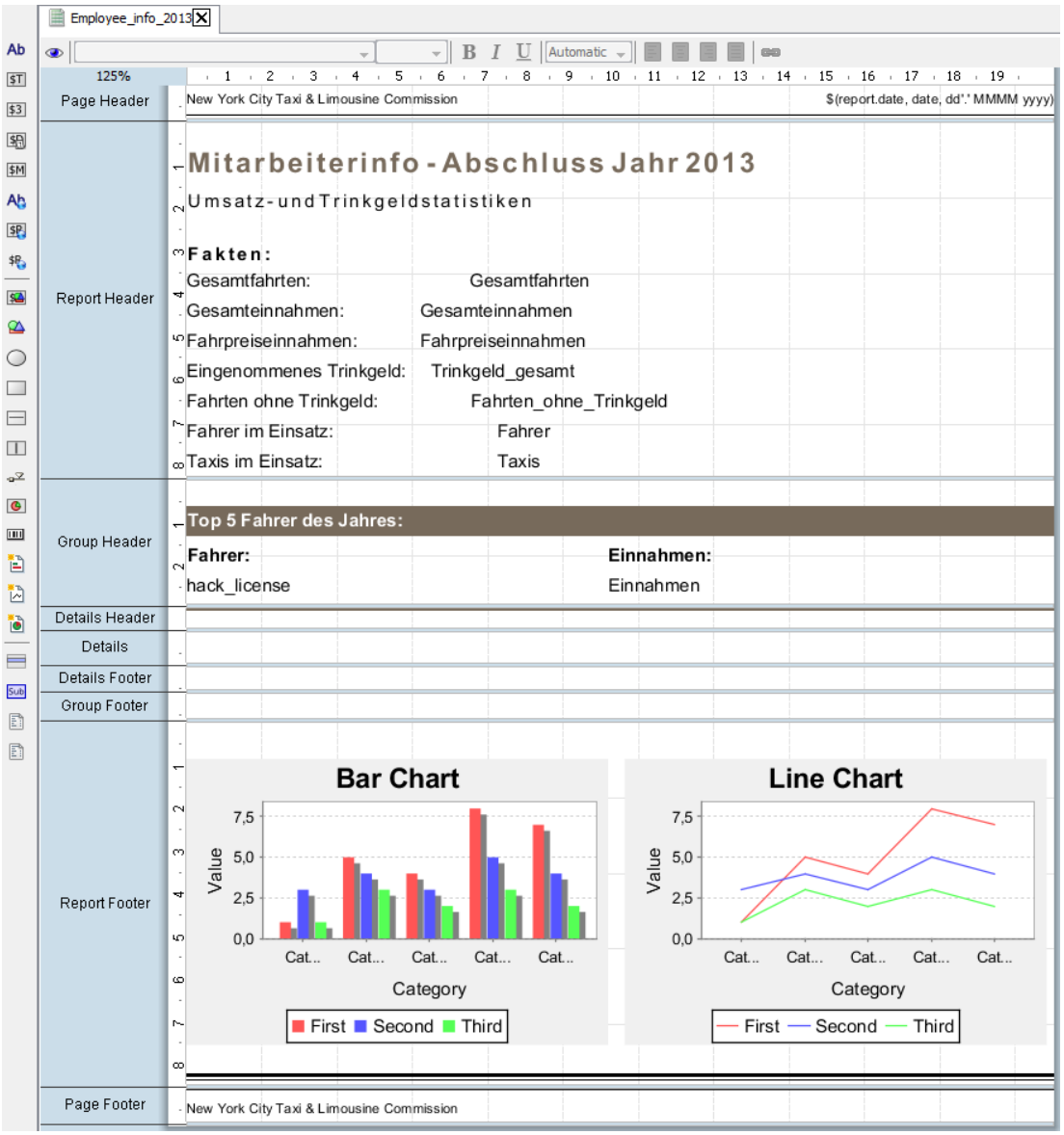
revenue_driver_total.xjpivot

Hack license	Kennzahlen: Transportations	Total amount
All Hack licenses	2.999.580	$ 44.326.693,30
3E7DE2A7DE12FE3F3E62AF403C8FEB94	690	$ 8.606,80
4FBAD760692D8C2B7A2ACE057B3B5F1B	646	$ 8.973,80
B3EF81B3356EBD377C83E11B1085F5D4	596	$ 7.518,00
946438FB059349E669781CE23362A176	585	$ 7.867,70
6067D592BE38133556DBD68DA07E2D24	582	$ 7.413,20
C547778100387B58D789079F2D0ECA17	573	$ 7.330,60
D07EB33A97693D232CEF070432EAA9B3	569	$ 7.075,50
4F50ACCC3F04E52FB0F88CC475C939B9	556	$ 6.935,10
518BCE0D10BDD35A6AF18DB1B87318BC	556	$ 7.902,00
C8D349DC61BCD170E1122B6503F8EE78	554	$ 8.193,50
DDA64A77E83A0FA134865A67788E9700	554	$ 6.739,00
9D5F4C4B01DAC9C86B17751BD8D23421	553	$ 8.255,70
CAFFA06F9610751C8B17EE3D6DDEA59E	553	$ 8.411,10
030C2C1ED4F1CEA429F4DA3DAE4F8B62	548	$ 8.130,60
75513D3470577109C3C0A5AE755C88AB	548	$ 8.052,10
9D11B6E767EE7FBED69D0ECE4FDBBF67	548	$ 7.555,60
3507FD423E811A560331455BB51EAC6C	547	$ 7.084,30
D18181CAC1222B3BFE18FDA6D8B43A2D	547	$ 6.477,60
C16A0343783175B25EA3FF2524EEFF86	544	$ 6.637,20
F6E7ADA84485264F2AF5AC88A8DDCBEA	543	$ 6.399,60
46198A4F28CD926181F388240895B204	531	$ 6.905,40
F5FFA4AA6DB399BB0838088FBC4CDF18	529	$ 6.749,80
0238047A5834FF6524C69E811F873D5A	528	$ 7.859,50

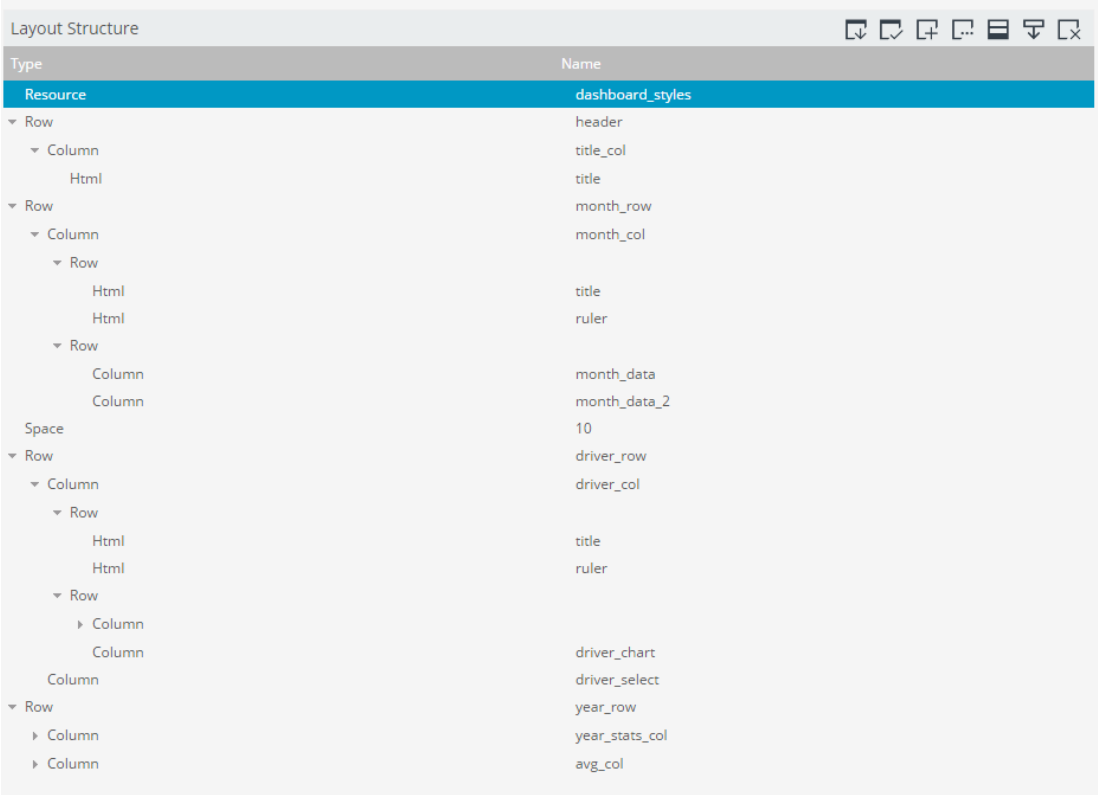
7.16 Analyse der umsatzstärksten Fahrer mittels Saiku Analytics



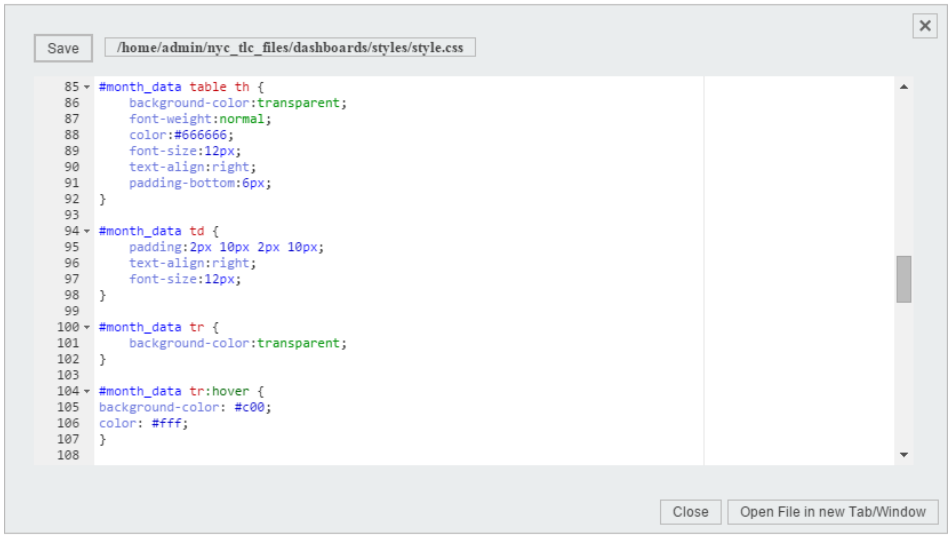
7.17 Entwurfsmodus der Personalanalyse



7.18 CDE – Erstellung eines Layouts



7.19 CDE – Erstellung eines Designs

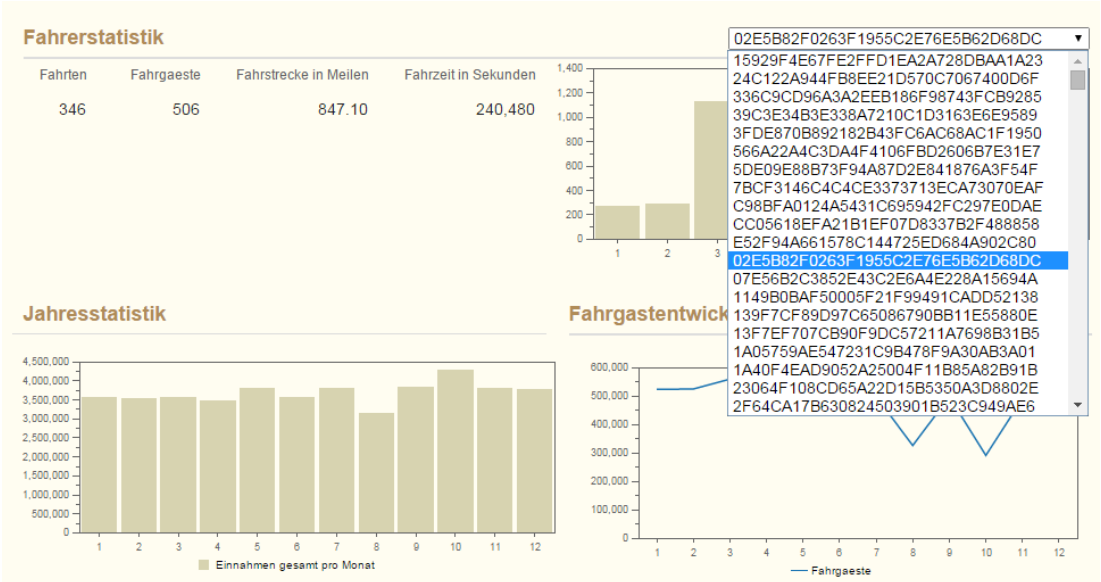


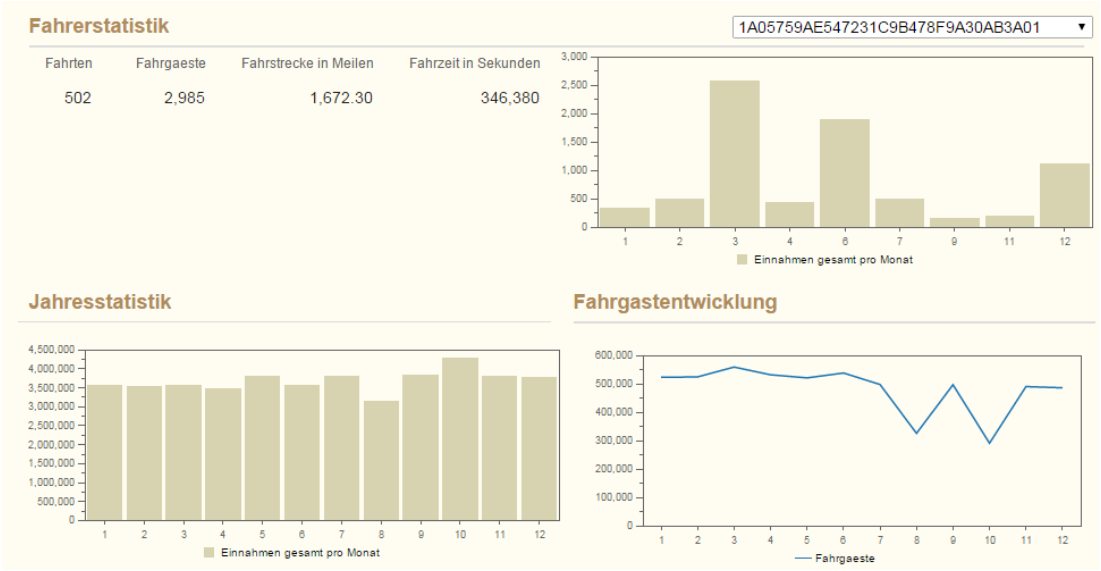
7.20 CDE – Angelegte Komponenten

Components

Type	Name
Group	Selects
Select Component	driver_picker
Group	Charts
CCC Bar Chart	year_chart_month
CCC Bar Chart	year_chart_month_driver
CCC Line Chart	year_passenger_chart
Group	Generic
Simple parameter	driver
Group	Others
table Component	driver_data_sql
table Component	month_data_sql
table Component	month_data_2_sql

7.21 CDE – Controlling Dashboard Ansichten




Fahrerstatistik
1A05759AE547231C9B478F9A30AB3A01
Fahrten
Fahrgaeste
Fahrstrecke in Meilen
Fahrzeit in Sekunden
502
2,985
1,672.30
346,380
Einnahmen gesamt pro Monat
Jahresstatistik
Einnahmen gesamt pro Monat
Fahrgastentwicklung
Fahrgaeste

Abkürzungsverzeichnis

BI ..... *Business Intelligence*
BSD ..... *Berkeley Software Distribution*

CDE ..... *Community Dashboard Editor*
CE ..... *Community Edition*
CRM ..... *Customer Relationship Management*
CSV ..... *Comma Separated Values*

EE ..... *Enterprise Edition*
EIS ..... *Executive Information System*
EMC ..... *EMC Corporation*
ETL ..... *Extract-Transform-Load*

GB ..... *Gigabyte*
GPL ..... *General Public Licence*

HOLAP ..... *Hybride Online Analytical Processing*
HTML ..... *Hypertext Markup Language*

IDC ..... *International Data Corporation*

LGPL ..... *Lesser General Public License*

MIT ..... *Massachusetts Institute of Technology*
MOLAP ..... *Multidimensional Online Analytical Processing*

OLAP ..... *Online Analytical Processing*
OS ..... *Open Source*
OSS ..... *Open Source Software*

PDF ..... *Portable Document Format*

ROLAP ..... *Relational Online Analytical Processing*
RTF ..... *Rich Text Format*

TLC ..... *New York City Taxi & Limousine Commission*

ZB ..... *Zettabyte*

Abbildungsverzeichnis

Tabellenverzeichnis

Literaturverzeichnis

Azevedo, Pedro, Gerhard Brosius, Stefan Dehnert, und Berthold Neumann. *Business Intelligence und Reporting mit Microsoft SQL Server 2008.* Microsoft, 2009.

Bachmann, Ronald, Guido Kemper, und Thomas Gerzer. *Big Data - Fluch oder Segen?: Unternehmen im Spiegel gesellschaftlichen Wandels.* mitp-Verlag, 2014.

Bange, Carsten. „Business Intelligence: Systeme und Anwendungen, Werkzeuge und Technologien für die Unternehmenssteuerung.“ *Controlling-Portal.* 2003. http://www.controllingportal.de/upload/iblock/b78/f9c03361d43d3f336a9534675deee0c0.pdf (Zugriff am 15. Januar 2015).

Bauer, Andreas, und Holger Günzel. *Data-Warehouse-Systeme: Architektur, Entwicklung, Anwendung.* 2013.

DATACOM Buchverlag GmbH. *ITWissen.* 2015. http://www.itwissen.info/definition/lexikon/Rotation-rotation.html (Zugriff am 17. Januar 2015).

Die Welt. 2015. http://www.welt.de/wirtschaft/webwelt/article118099520/Datenvolumen-verdoppelt-sich-alle-zwei-Jahre.html (Zugriff am 20. Januar 2015).

Gabriel, Roland. *Enzyklopädie der Wirtschaftsinformatik.* 13. September 2013. http://www.enzyklopaedie-der-wirtschaftsinformatik.de/wi-enzyklopaedie/lexikon/uebergreifendes/Kontext-und-Grundlagen/Informationssystem/Fuhrungsinformationssystem (Zugriff am 12. Dezember 2014).

„Google Charts.“ 2015. https://developers.google.com/chart/interactive/docs/gallery/geomap (Zugriff am 22. März 2015).

Güclü, Ilkem, und Stefan Müller. „Business Intelligence mit Open Source: Vergleich der BI-Lösungen Jaspersoft, Jedox Palo und Pentaho.“ *t3n Magazin Nr. 19*, März 2010.

Hansen, Robert, und Gustaf Neumann. *Wirtschaftsinformatik 1 Grundlagen und Anwendungen.* Lucius & Lucius Verlagsgesellschaft mbH, 2009.

ifrOSS. *Institut für Rechtsfragen der Freien und Open Source Software.* 2015. http://www.ifross.org/welches-sind-wichtigsten-open-source-lizenzen-und-welchem-lizenztyp-gehoeren-sie (Zugriff am 22. Januar 2015).

it-novum GmbH. „Controlling-Portal.“ *www.controllingportal.de.* 2009. http://www.controllingportal.de/upload/iblock/b96/01fc309d681b51b36a9523596098fae4.pdf (Zugriff am 17. Dezember 2014).

Jagadish, H. V., et al. „Communications of the Association for Computing Machinery.“ *Big data and its technical challenges.* 2014. http://dl.acm.org/citation.cfm?id=2622628.2611567&coll=DL&dl=GUIDE&CFID=635430525&CFTOKEN=36102858 (Zugriff am 08. März 2015).

Kemper, Hans-Georg, Henning Baars, und Walid Mehanna. *Business Intelligence – Grundlagen und praktische Anwendungen.* Vieweg+Teubner Verlag, 2010.

Klein, Dominik, Phuoc Tran-Gia, und Matthias Hartmann. *Gesellschaft für Informatik.* 2013. http://www.gi.de/service/informatiklexikon/detailansicht/article/big-data.html (Zugriff am 08. März 2015).

Laney, Doug. „Gartner.“ *Deja VVVu: Others Claiming Gartner's Construct for Big Data.* 2012. http://blogs.gartner.com/doug-laney/deja-vvvue-others-claiming-gartners-volume-velocity-variety-construct-for-big-data (Zugriff am 04. März 2015).

LfDI RLP. *YOUNG DATA.* 2015. http://www.youngdata.de/internet/internet-der-dinge (Zugriff am 24. Januar 2015).

Manyika, James, et al. „Big data: The next frontier for innovation, competition,and productivity.“ *McKinsey Global Institute.* 2011. http://www.mckinsey.com/~/media/McKinsey/dotcom/Insights%20and%20pubs/MGI/Research/Technology%20and%20Innovation/Big%20Data/MGI_big_data_full_report.ashx (Zugriff am 24. Mai 2015).

Open Source Initiative. *Open Source Initiative.* 2014b. http://opensource.org/licenses/category (Zugriff am 13. Dezember 2014).

—. *Open Source Initiative.* 2014a. http://opensource.org/docs/osd (Zugriff am 4. Dezember 2014).

„OpenStreetMap.“ *Genauigkeit von Koordinaten.* 2015. http://wiki.openstreetmap.org/wiki/DE:Genauigkeit_von_Koordinaten (Zugriff am 22. März 2015).

Oppermann, Reinhard, und Harald Reiterer. *Software-ergonomische Evaluation.* 1994.

Pentaho Corporation. *Pentaho.* 2014. http://www.pentaho.de/about/story (Zugriff am 17. Dezember 2014).

Pentaho DI Documentation. *Latest Pentaho Data Integration (aka Kettle) Documentation.* 2014. http://wiki.pentaho.com/display/EAI/Latest+Pentaho+Data+Integration+%28aka+Kettle%29+Documentation (Zugriff am 19. Dezember 2014).

Pentaho DM Documentation. „Pentaho Data Mining Community Documentation.“ 2014. http://wiki.pentaho.com/display/DATAMINING/Pentaho+Data+Mining+Community+Documentation (Zugriff am 20. Dezember 2014).

Pentaho Mondrian. *Mondrian.* 2014. http://community.pentaho.com/projects/mondrian (Zugriff am 20. Dezember 2014).

Pentaho Reporting. *Pentaho.* 2014. http://community.pentaho.com/projects/reporting (Zugriff am 19. Dezember 2014).

Rödl Dynamics AG. *intelligence.de.* 8. Januar 2014. http://www.intelligence.de/news/technologische-umsetzung-von-business-intelligence-systemen.html (Zugriff am 10. Dezember 2014).

Rohling, Gitta. *Siemens - Pictures of the Future.* 2014. http://www.siemens.com/innovation/de/home/pictures-of-the-future/digitalisierung-und-software/internet-of-things-fakten-und-prognosen.html (Zugriff am 25. Januar 2015).

Schneider, Wolfgang. *Ergonomische Gestaltung von Benutzungsschnittstellen: Kommentar zur Grundsatznorm DIN EN ISO 9241-110.* Herausgeber: DIN e.V. Beuth, 2008.

Statista GmbH. *Statista.* 2015. http://de.statista.com/statistik/daten/studie/267974/umfrage/prognose-zum-weltweit-generierten-datenvolumen (Zugriff am 21. Januar 2015).

TLC. *New York City Taxi and Limousine Commission.* 05. Januar 2014. http://www.nyc.gov/html/tlc/html/about/about.shtml (Zugriff am 05. Januar 2014).

Wang, Richard Y., und Diane M. Strong. „University of Washington." *www.courses.washington.edu.* 1996. http://courses.washington.edu/geog482/resource/14_Beyond_Accuracy.pdf (Zugriff am 27. Januar 2015).

Webdetails. *Webdetails.* 2014. http://www.webdetails.pt/ctools/cde/#section=overview (Zugriff am 20. Dezember 2014).

Versicherung über Selbstständigkeit

Hiermit versichere ich, dass ich die vorliegende Arbeit ohne fremde Hilfe selbstständig verfasst und nur die angegebenen Hilfsmittel benutzt habe.

Hamburg, den _____ _____