IBM BigInsights Workshop

Jonas Freiknecht, Information Architect

jonas.freiknecht@de.ibm.com

Einführung

BigInsights ist IBMs Big-Data-Plattform, die auf Basis von Apache Hadoop die Verarbeitung großer Datenmengen erlaub. Sie besteht zum einen aus den bekannten opensource Komponenten wie Apache Hive, HBase, Flume, Sqoop, Zookeeper etc. und wird zum anderen durch einige Erweiterungen angereichert, die die Arbeit mit großen, unstrukturierten Daten erleichtern. Dazu gehört etwa die Sprache AQL (Annotation Query Language), um Fließtext zu analysieren und daraus geschäftskritische Informationen zu gewinnen, BigSQL, um auf großen Datenmengen über SQL Abfragen auszuführen oder BigSheets, um in einer excel-ähnlichen Umgebung Daten tabellarisch darzustellen, zu filtern, aufzubereiten und zu visualisieren.

In diesem Workshop soll gezeigt werden, wie Sie BigInsights bedienen und dessen Komponenten einsetzen, um für Ihre Ansprüche und Usecases die passende Methode zu finden, diese umzusetzen. Es wird mit einer Übersicht über die Web-Console begonnen, die Ihnen beim Navigieren in BigInsights helfen soll, um die gewünschten Funktionen schnell und einfach zu finden. Im zweiten Teil geht es darum, Daten von einem lokalen System oder einem FTP-Server nach Hadoop zu transportieren, um diese dort analysieren zu können. Dabei betrachten wir einerseits den webbasierten HDFS-Explorer in BigInsights und andererseits die Anwendungen zum Dateiimport, die direkt über die Web-Console gestartet werden können. Im dritten Schritt wird BigSheets vorgestellt, indem Sie einige einfache Daten in ein Workbook laden, die Daten aufbereiten und am Ende visualisieren. Es schließt sich im vierten Schritt eine Erklärung zu BigSQL an, in der es darum geht eine Verbindung zum BigSQL-Server herzustellen, Tabellen zu laden, Daten zu importieren und damit zu arbeiten. Als Königsdisziplin soll dann im fünften Kapitel erklärt werden, wie Sie eigene Anwendungen auf Basis von Java, MapReduce, Oozie und AQL entwickeln und auf BigInsights deployen können.

Voraussetzungen

Sie sollten ein grundliegendes Verständnis davon mitbringen, was Big-Data ist und was es ausmacht. Desweiteren ist es hilfreich die Architektur und Funktionsweise von Hadoop und dessen einzelne Komponenten MapReduce, YARN und das HDFS zu kennen. Auf technischer Seite sollten Sie Zugriff auf eine Virtual Machine mit einem vorinstallierten BigInsights 3.0 haben[1].

[1] Sollte das nicht der Fall sein, können Sie ein solches aufsetzen, indem Sie diesem Tutorial folgen: https://www.youtube.com/watch?v=_yVYFbBR0Do

Lektion 1 - Übersicht über BigInsights

Melden Sie sich zu Beginn mit den Benutzerdaten an der BigInsights-Console an. Sie erreichen diese über folgenden Link:

http://localhost:8080/

Der Hostname kann natürlich variieren, Sie sollten hier denjenigen angeben, der auf den Masterknoten Ihres BigInsights-Clusters verweist. Wenn Sie lediglich, wie für die Demozwecke vorgesehen, einen Knoten in einem sogenannten pseudo-distributed Cluster verwenden, dann geben Sie den Hostnamen oder die IP dieses einen Knoten an.

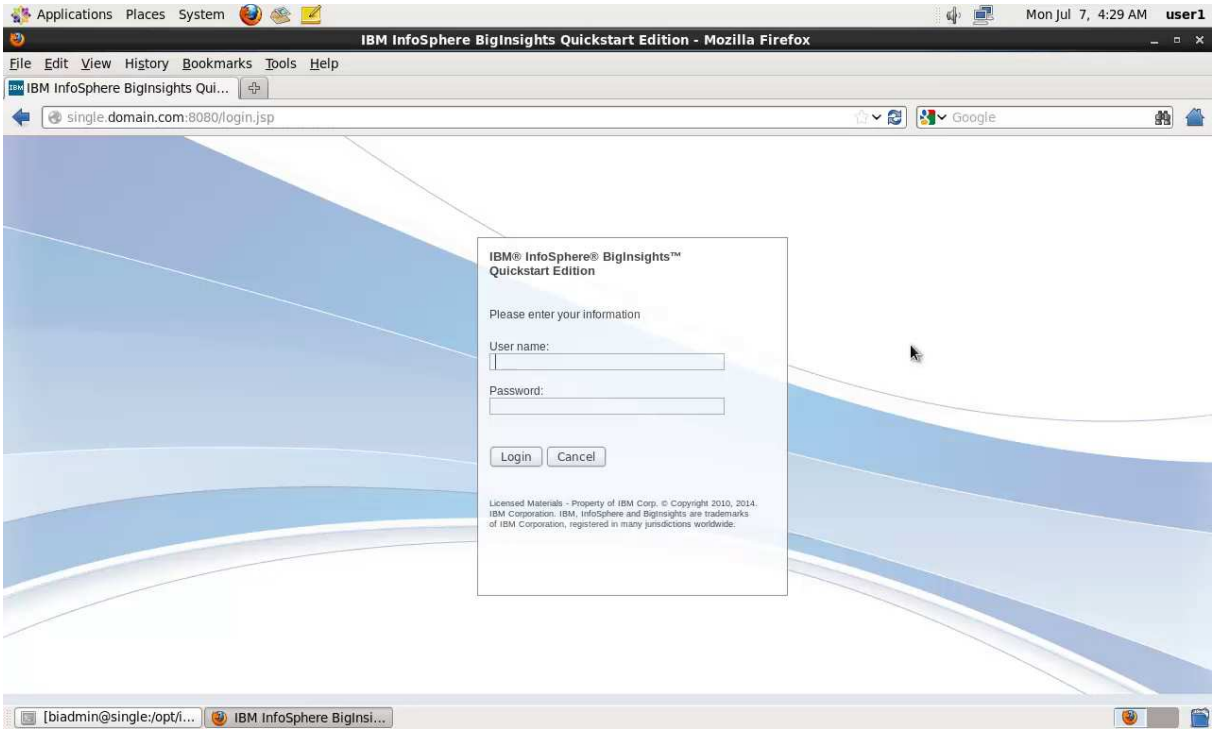


Abbildung 1: Anmelden an der BigInsights-Console

Sie starten üblicherweise im Welcome-Screen und werden dort unter anderem die Gruppierungen *Tasks*, *Quick Links* und *Learn More* sehen.

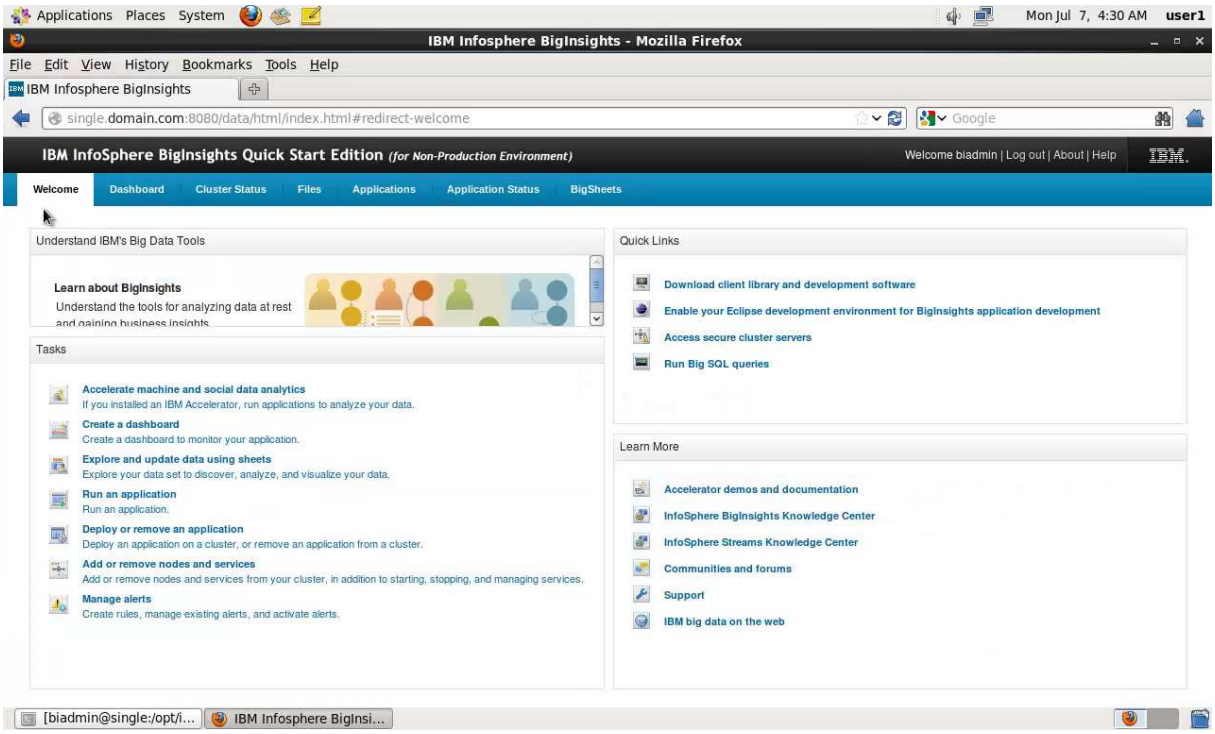


Abbildung 2: Der Welcome-Screen

Unter **Tasks** werden Sie zu den Sichten umgeleitet, die Sie benötigen, um die üblichen Aufgaben in BigInsights zu erledigen. Darunter etwa das Anlegen eines neuen Dashboards zur Datenvisualisierung oder dem Deployment und dem Ausführen einer neuen Anwendung. **Quick Links** richtet sich ein wenig mehr an Entwickler und erlaubt es etwa, ein Eclipse so einzurichten, um damit Anwendungen für BigInsights entwickeln zu können oder aber es wird gezeigt, wo man mit BigSQL einfache Abfragen ausführt und generell mit SQL auf seinen großen Datenmengen arbeiten kann. Die letzte Gruppe, **Learn More**, bietet, wie der Name schon vermuten lässt, einige weiterführende Informationen z.B. über das *Knowledge Center* (Dem Nachfolger des Information Centers).

Wechseln Sie nun in den Tab **Dashboard** in der oberen Menüzeile. Dort gibt es noch nicht sehr viel zu sehen, bis auf eine Combobox mit den drei standardmäßig vorhandenen Dashboards, die Auskunft über den Clusterstatus geben oder Informationen über ausgeführte, verteilte Anwendungen im Detail vorhalten.

Wundern Sie sich nicht, wenn Sie zu Beginn in jedem der Dashboards eine Error-Meldung sehen. Diese weist lediglich darauf hin, dass noch keine Daten existieren, die hier dargestellt werden können, da wir das System ja zuvor ganz frisch aufgesetzt haben.

Im nächsten Tab, **Cluster Status**, finden Sie den jeweiligen Zustand der Komponenten von BigInsights vor. Oben wird neben dem Bezeichner Nodes aufgeführt, wie viele Knoten zurzeit im Cluster aktiv sind. Desweiteren finden Sie die Bestandteile von Hadoop vor, das HDFS und MapReduce, sowie andere Programme aus dem Apache Stack, z.B. Hive, Base, Oozie und Zookeeper.

Dass die Komponente **Monitoring** nicht läuft ist kein Fehler sondern gewollt. Sie wird im Cluster mit nur einem Knoten per Default deaktiviert, um Ressourcen zu sparen. Wählen Sie diese auf Wunsch an und klicken Sie im rechten Fenster auf *Start*, um sie zu aktivieren.

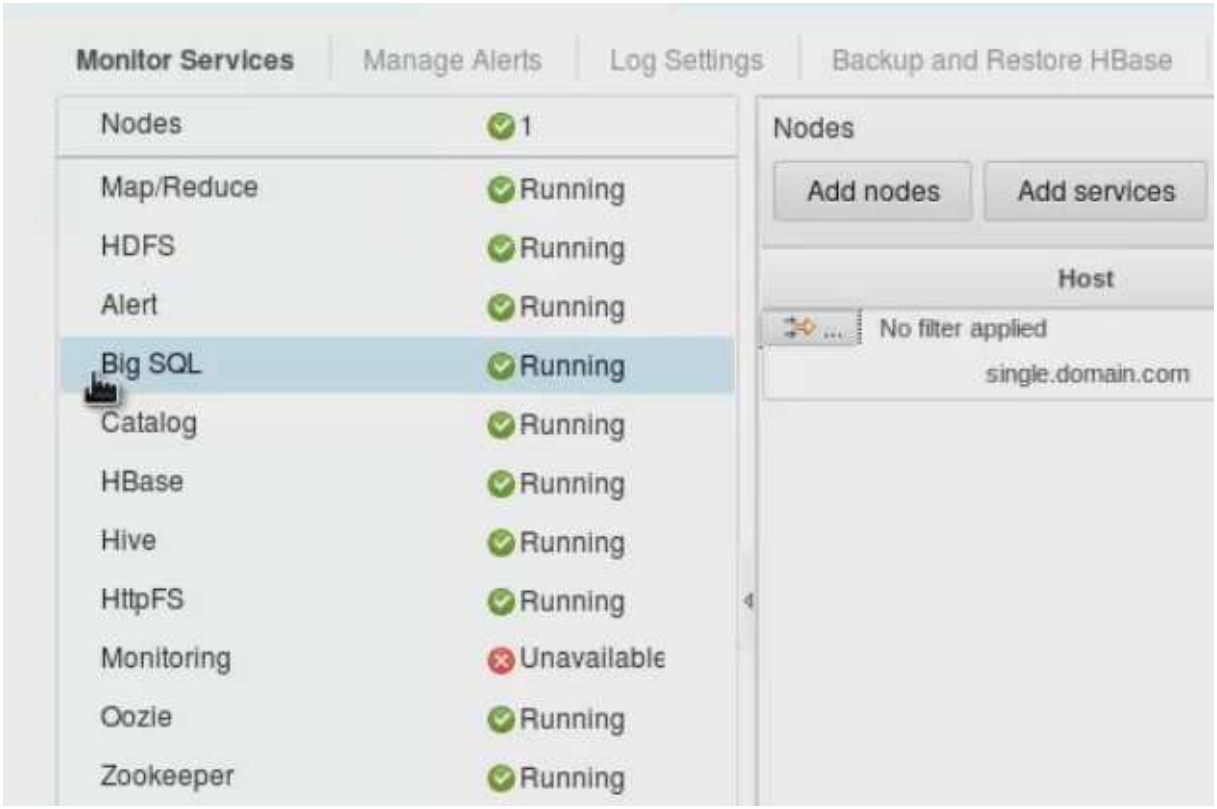


Abbildung 3: Zustand der Komponenten im Cluster

Der Tab **Files** birgt eine der Sichten, in der Sie sich sicherlich zu Anfang recht häufig aufhalten werden. Darin können Sie ähnlich einem gängigen Dateiexplorer auf die Daten des Hadoop Distributed File Systems (*HDFS*) zugreifen und Dateien erstellen, manipulieren und löschen.

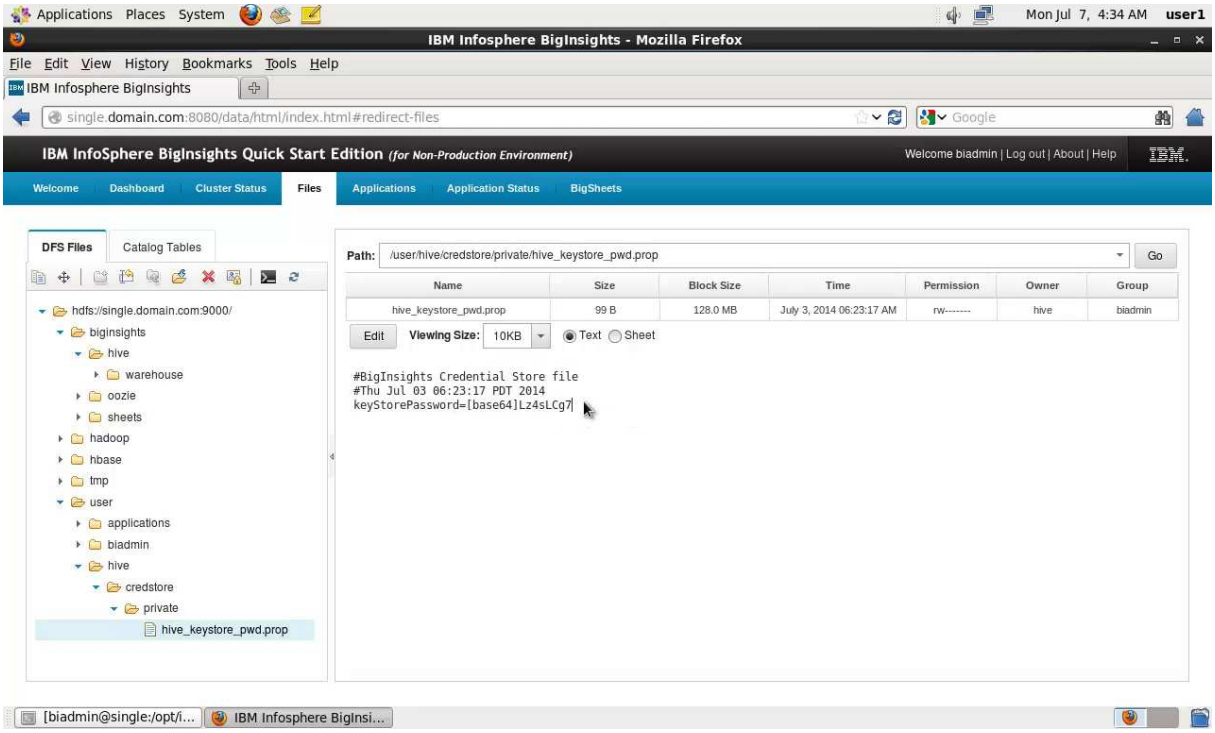


Abbildung 4: Zugriff auf das HDFS

Sie finden im linken Fenster unter dem Reiter *DFS Files* einige Optionen, um mit den Dateien darin zu arbeiten. Diese sind von links nach rechts:

- Kopieren
- Verschieben
- Neuen Ordner anlegen
- Umbenennen
- Vom lokalen Dateisystem hochladen
- Herunterladen
- Löschen
- Berechtigungen ändern
- HDFS-Konsolenbefehl ausführen
- Ansicht aktualisieren

Diese Optionen sollten weitestgehend selbsterklärend sein. Wenn Sie bereits mit Hadoop gearbeitet haben und sich mit den üblichen *Shell-Commands* auskennen, können Sie über die vorletzte Option auch komplexere Befehle direkt in Textform eingeben. Da die Webansicht jedoch alle wichtigen Funktionalitäten (bis auf ein *tail -f*) abbildet, wollen wir nicht weiter auf diesen Punkt eingehen.

Unter der Leiste mit Befehlen finden Sie die Baumansicht. Das Hauptelement, hdfs://single.domain.com:9000, zeigt auch gleichzeitig den Stamm der URL, die Sie benötigen, wenn Sie z.B. aus anderen Anwendungen auf dieses HDFS zugreifen möchten. Klicken Sie nun in der Baumansicht eine Datei an, dann wird diese im rechten Fenster in einer Art Vorschau gezeigt. Große Dateien werden nie zur Gänze geladen sondern immer nur ein Auszug dessen Größe Sie über die Combobox *Viewing Size* spezifizieren können. Für Ordner und Dateien finden Sie ebenso Informationen über das selektierte Element wie Größe oder Unterordner in der Tabelle über der Dateivorschau.

Der Tab **Applications** ist in drei Unterfenster aufgeteilt, **Run**, **Manage** und **Link**. **Run** lässt uns die bereits deployten Anwendungen ausführen, wie in Abbildung 5 zu sehen. *Manage* hingegen verwaltet alle verfügbaren Anwendung. Achtung, nicht jede Anwendung, die auf BigInsights verfügbar ist, ist auch gleichzeitig deployed. Nur Anwendung, die über den Reiter Manage deployed wurden, können auch verwendet werden.

Wählen Sie **Manage** aus, so sehen Sie verschiedene Anwendungskategorien, z.B. *Import*, *Export*, *SQL* oder *Web*. Jede Anwendung, die Sie erstellen oder die mit BigInsights ausgeliefert wird, kann mehreren Kategorien zugeordnet werden. Das hilft dabei, für jede Anforderung schnell die richtige Anwendung zu finden. Wenn Sie eine der Anwendungen verwenden möchten, selektieren Sie diese und klicken Sie im rechten Fenster auf *Deploy* und im sich öffnenden Fenster erneut auf *Deploy*. Nun wird die Anwendung erscheinen, wenn Sie erneut auf den Tab *Run* wechseln. Im nächsten Kapitel werden wir sehen, wie eine Anwendung verwendet wird.

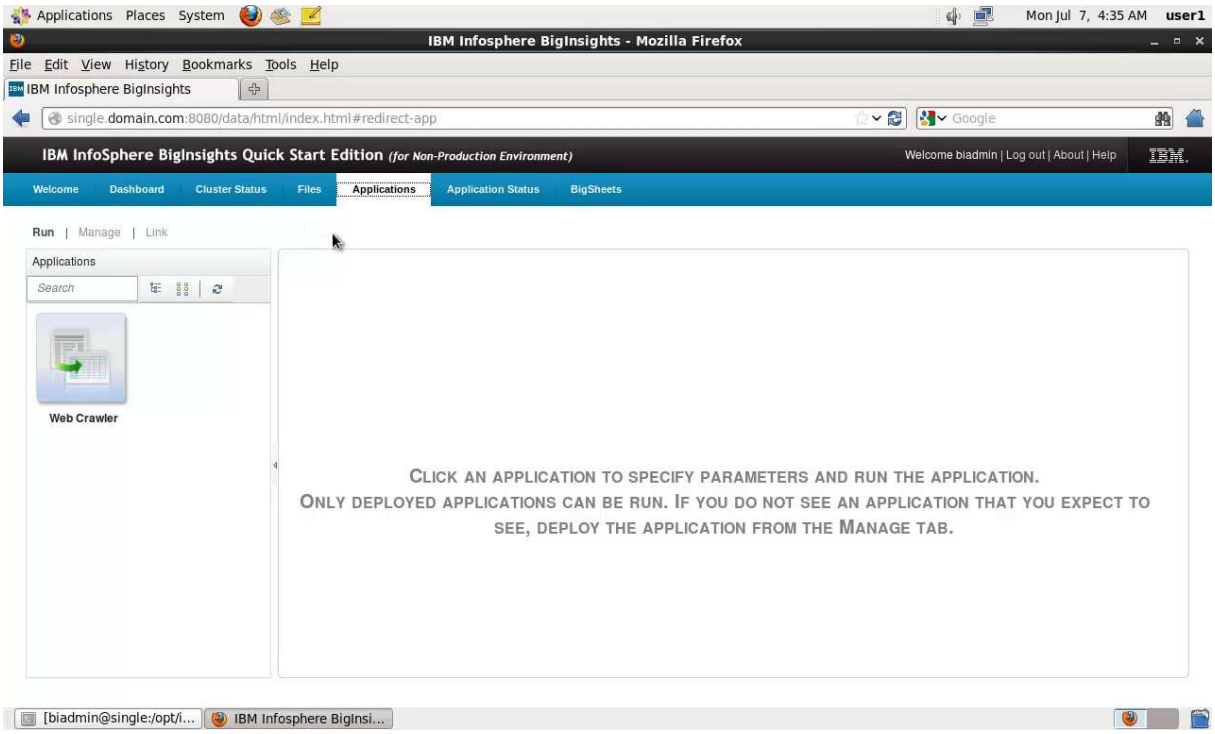


Abbildung 5: Übersicht über die deployten Anwendungen

Unter dem Reiter **Link** können Sie mehrere Anwendungen verknüpfen. Wenn Sie z.B. den Web Crawler nutzen, um Internetseiten herunterzuladen könnte ein Folgeszenario sein, diese Seiten mit einem Word Count hinsichtlich bestimmter, auftauchender Begriffe zu analysieren. Diesen Sie dazu die Anwendung *Web Crawler* (Sie müssen diese vorher deployen) auf das gestrichelte Kästchen und danach die Anwendung *Word Count* auf das neue gestrichelte Kästchen. Nun müssen Sie in diesem Usecase die Anwendungen so konfigurieren, dass das Ausgabeverzeichnis des *Web Crawlers* das Eingabeverzeichnis der *Word-Count-Anwendung* ist und die Anwendungen werden dann beim Ausführen der Verlinkung nacheinander gestartet. Zuvor müssen Sie diese jedoch noch als neue Anwendung speichern, die Konsole führt Sie auf Wunsch durch diesen Vorgang.

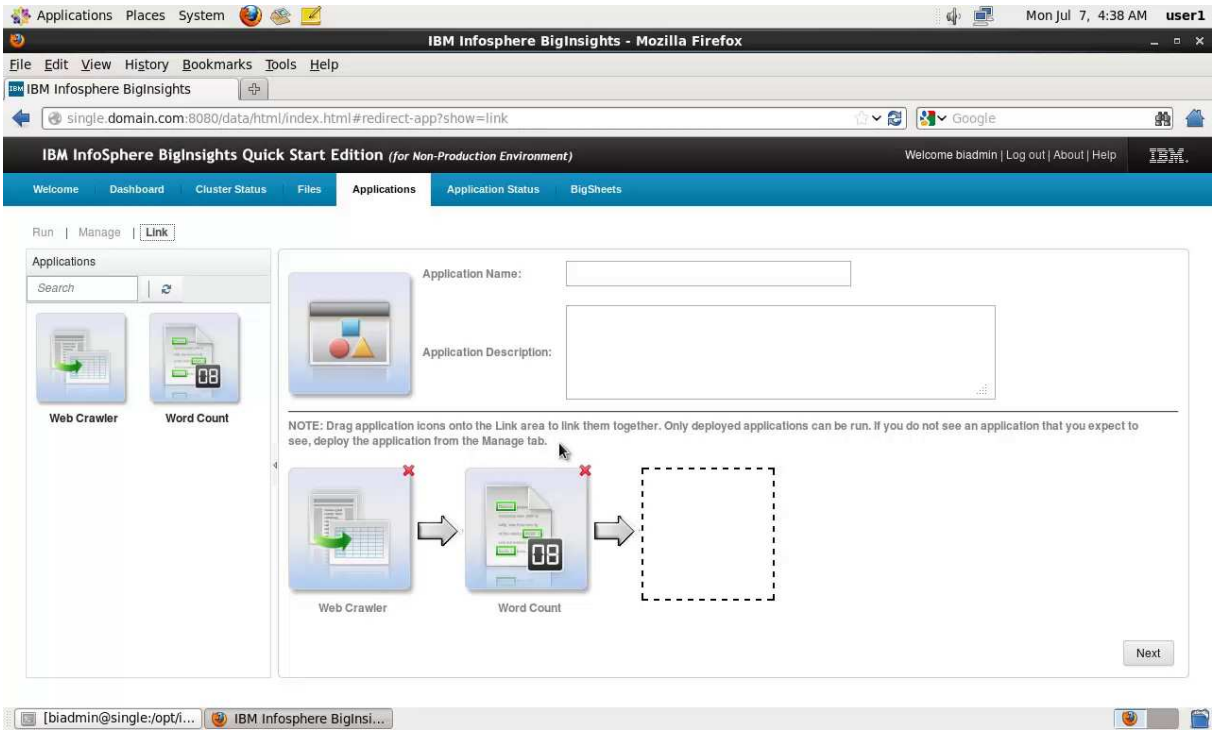


Abbildung 6: Verlinken zweier Anwendungen

Wenn Sie dann eine oder mehrere Anwendungen ausgeführt haben, können Sie deren Status unter dem Tab **Application Status** beobachten. Darin wird gezeigt welche Anwendungen gerade laufen und welche ihre Arbeit mit welchem Ergebnis verrichtet haben.

Der Reiter **BigSheets** ist der letzte und dazu auch noch einer der mächtigsten. Hier bietet Ihnen BigInsights eine excel-ähnliche Umgebung, um Daten im HDFS tabellarisch und in Form von Diagrammen darzustellen. Da BigSheets Bestandteil eines ganzen Kapitels ist, soll hier noch nicht weiter darauf eingegangen werden.

Lektion 2 - Datenimport

Um nun eine Datenbasis zu haben, mit der wir arbeiten können, wenn wir uns BigSQL und BigSheets anschauen und später noch eine eigene Anwendung implementieren werden, müssen wir zuerst ein paar Daten in das HDFS unter BigInsights importieren.

Eine, die einfachste, Möglichkeit haben wir bereits kennengelernt. Über den Tab Files haben wir Zugriff auf den HDFS-Explorer, der es uns über den Upload-Button ermöglicht, Dateien von dem Rechner von dem aus wir auf die Web-Oberfläche zugreifen, in das HDFS zu laden. Ich möchte im Root-Verzeichnis nun folgende Ordnerstruktur anlegen *apps* → *customeranalysis* → *input*. Verwenden Sie dazu die Funktion, um neue Ordner zu erzeugen und selektieren Sie immer das Vaterelement, um darin einen Unterordner zu erstellen. Ich wähle die Ordnernamen beliebig. Der Ordner *apps* soll dabei alle Daten für unsere eigenen Datenanalysen vorhalten. Der Unterordner *customeranalysis* soll unseren derzeitigen Usecase darstellen und *input* hält dafür die Eingabedaten vor.

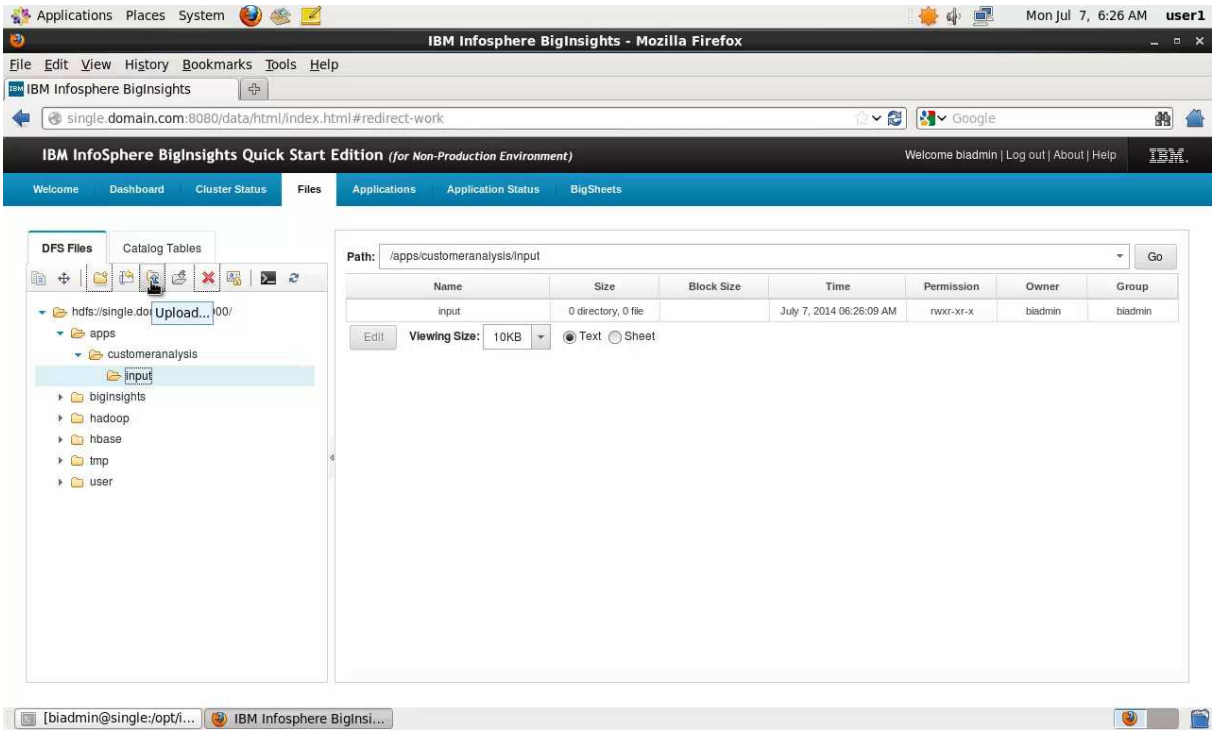


Abbildung 7: Anlegen einer Ordnerstruktur

Klicken Sie nun wie in Abbildung 7 zu sehen auf den Ordner *input* und wählen Sie über das Menü die Funktion Upload. Es öffnet sich ein neues Fenster, über das Sie eine lokale Datei auswählen können. Wählen Sie die Datei *customers.csv*, die mit diesem Tutorial ausgeliefert wird und klicken Sie auf *OK*. In der Liste *Files to Upload* können Sie nun noch mehr Dateien für den Upload auswählen, uns soll jedoch diese eine genügen.

Im rechten Fenster sollte nach dem Hochladen direkt eine Vorschau der Datei zu sehen sein.

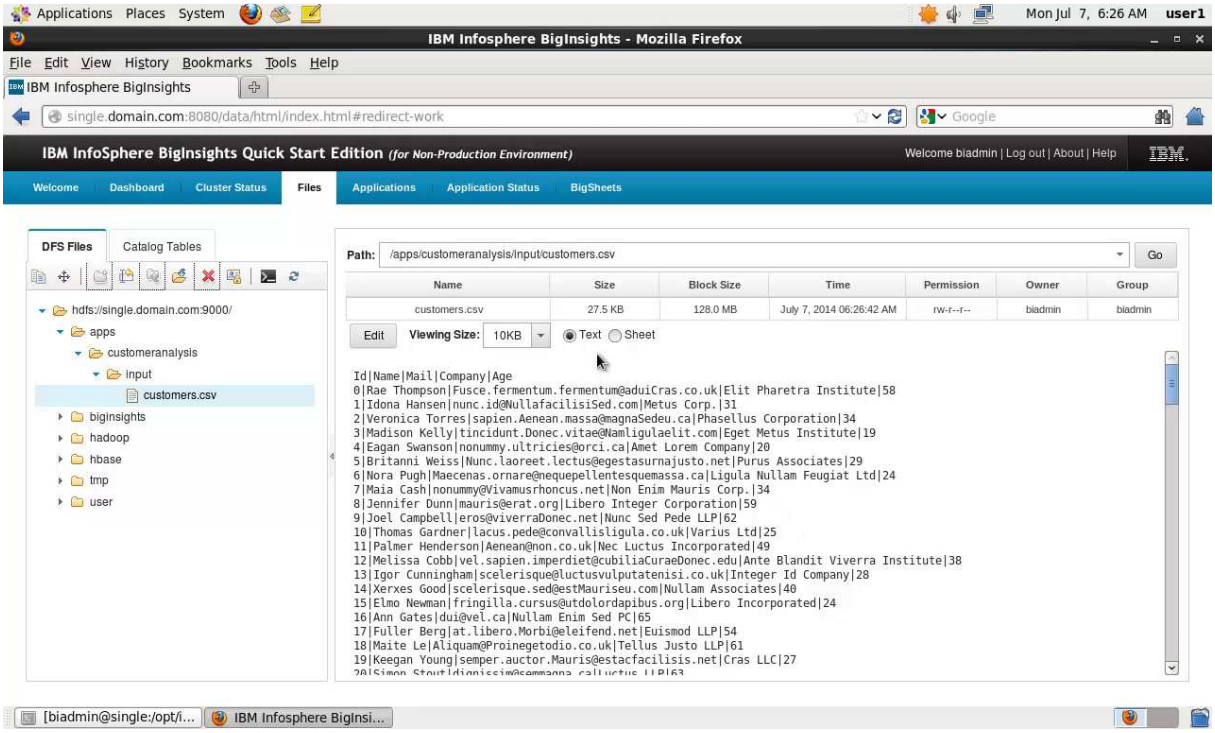


Abbildung 8: Vorschau der hochgeladenen Datei

Sehr gut, nun haben wir eine CSV-Datei, mit der wir später die ersten Datenanalysen ausführen können.

Ein anderer Weg, um Daten in das HDFS zu bekommen, führt über den Tab **Applications**, in dem es eine ganz eigene Kategorie namens **Import** gibt.

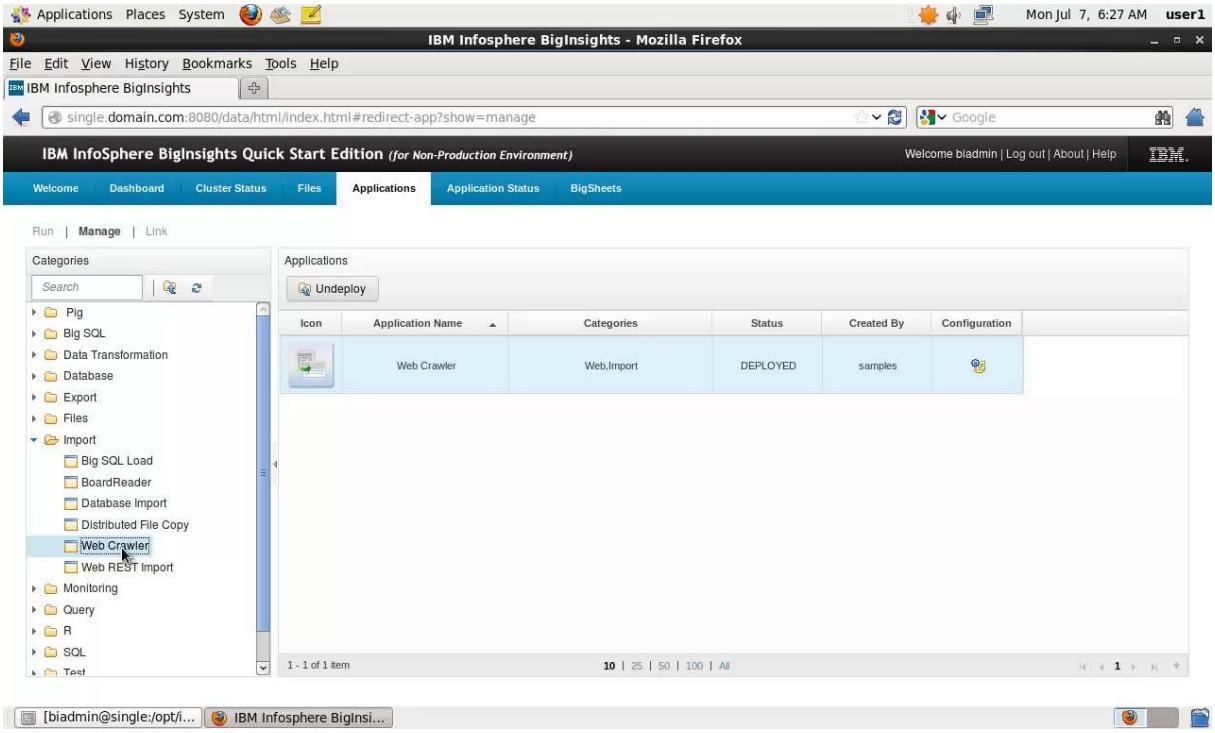


Abbildung 9: Anwendungen für den Datenimport

Wir wollen die Anwendung **Distributed File Copy** verwenden, um nun Daten von einem FTP-Server in das HDFS zu laden. Wählen Sie diese Anwendung aus und deployen Sie sie, falls das noch nicht im Vorfeld geschehen sein sollte (Sie müssen sich dazu im Tab *Manage* befinden). Im Anschluss sollten Sie die Anwendung unter *Run* wiederfinden.

Konfigurieren Sie diese entsprechend der folgenden Abbildung. Dazu müssen Sie einen lokalen FTP-Server installiert haben, der einen Benutzer mit Namen *user1* und Passwort *user1* zulässt. Falls Sie einen FTP-Server verwenden möchten, der woanders läuft, ändern Sie einfach die Zugangsdaten.

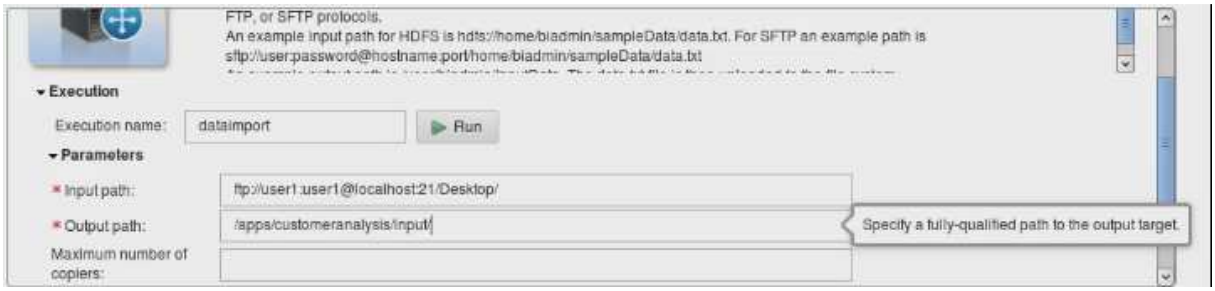


Abbildung 10: Konfiguration der Datenimportanwendung

Die Parameter der Anwendung sind obligatorisch, wenn sie mit einem roten Sternchen gekennzeichnet sind. Der Inputpfad ist die übliche URL für den Zugriff auf einen FTP-Server, beginnen mit dem Protokoll, dem Benutzernamen und dem Passwort, gefolgt von Host, Port und Ordner. Wir wollen also alle Daten von dem FTP laden, die sich im Verzeichnis *Desktop* befinden. Dieses

Verzeichnis voll dann, wie in *Output path* spezifiziert, in den zuvor erstellten Ordner */apps/customeranalysis/input* abgelegt werden. Optional können Sie der Ausführung einen Namen geben, ich habe diese, wie oben zu sehen, *dataimport* genannt. Indem Sie den Prozess benennen, finden Sie ihn in der History leichter wieder und können so nachschauen, wie sie ihn z.B. bei der letzten Ausführungen konfiguriert haben, denn alle Parameter werden in der History gespeichert. Klicken Sie anschließend auf **Run** und der Import beginnt. Nach etwa 30 Sekunden sollten Sie im unteren Teil des Fensters eine Erfolgsmeldung sehen.

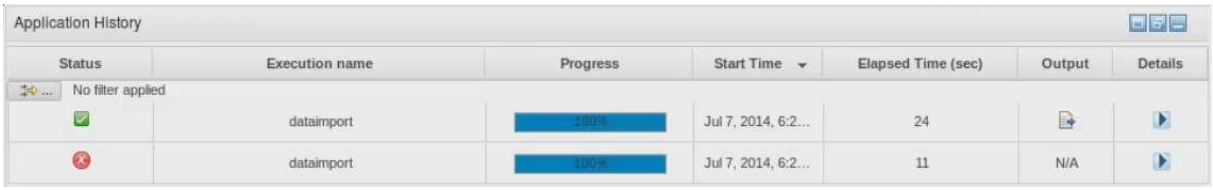


Abbildung 11: History der Importanwendung

Ein grünes Häkchen markiert einen erfolgreichen Import, ein rotes Kreuz einen fehlgeschlagenen. Der Import kann z.B. fehlschlagen, wenn die Zugangsdaten zu dem FTP-Server inkorrekt sind, über den Pfeil ganz rechts unter *Details* erhalten Sie weitere Informationen darüber, warum der Prozess nicht erfolgreich beendet wurde.

Navigieren Sie nun wieder in den Tag **Files**, selektieren Sie den Ordner *input* und klicken Sie auf die beiden blauen Pfeile rechts über den Baumstruktur, um die Ansicht zu aktualisieren.

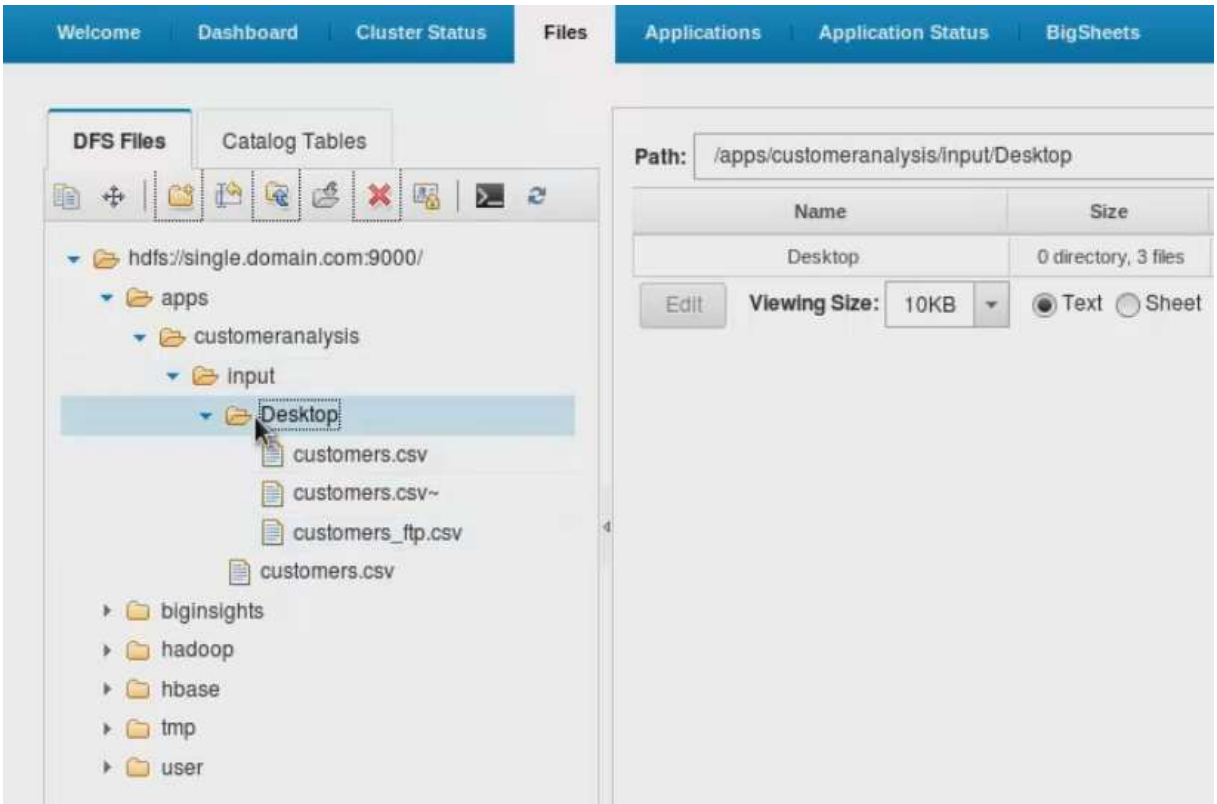


Abbildung 12: Importierte Daten im HDFS

Die Importanwendung hat nun den gesamten Ordner Desktop vom FTP in das HDFS kopiert. Für uns ist nur die Datei *customers_ftp.csv* interessant, also markieren wir diese, klicken auf den

Verschiebepfeil und wählen als Ziel den Ordner *input*. Anschließend löschen wir im HDFS den gesamten Ordner Desktop wieder über das rote Kreuz (Ordner muss selektiert sein).

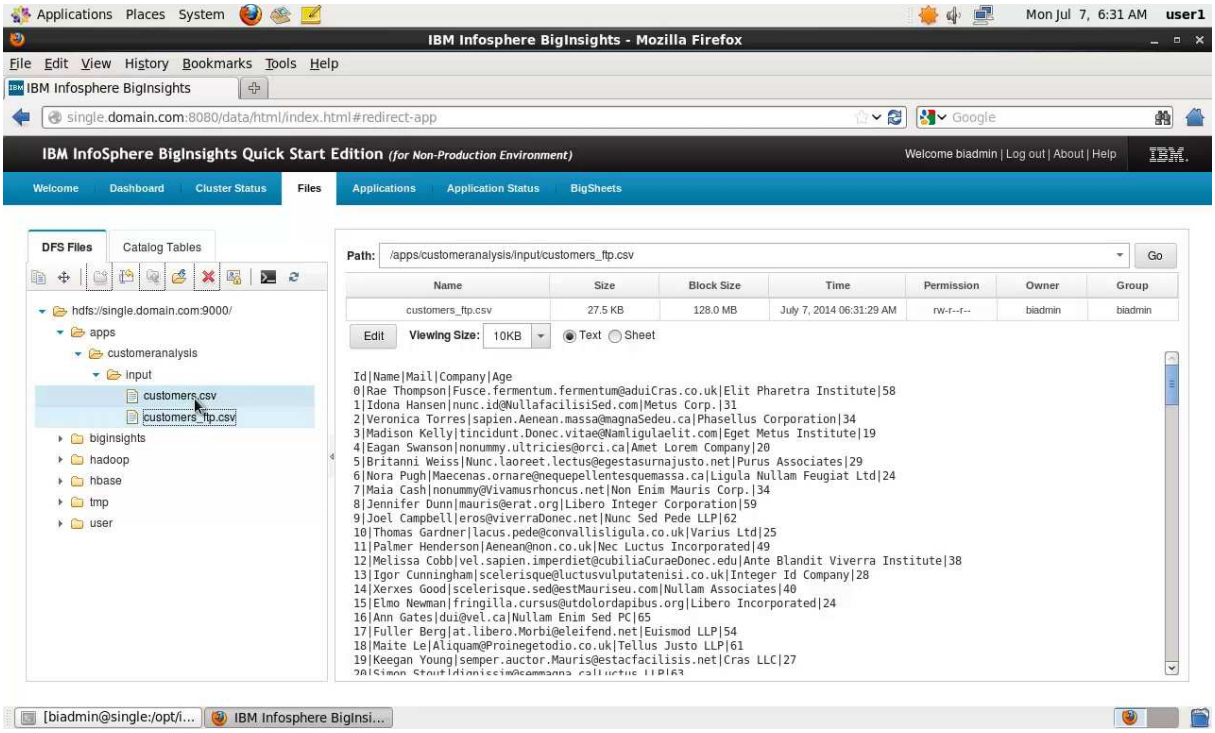


Abbildung 13: Die zwei importierten Dateien im HDFS

Nun haben wir gesehen, wie einfach es ist, Dateien in das HDFS zu importieren. Im folgenden Kapitel wollen wir diese nun mithilfe von BigSheets analysieren und visualisieren.

Lektion 3 - Datenanalyse mit BigSheets

In dieser Lektion soll nun gezeigt werden, wir die eben importierten Daten über *BigSheets* analysiert und visualisiert werden können. Aus dem Tab *Welcome* wechseln wir in den Tab *BigSheets*.

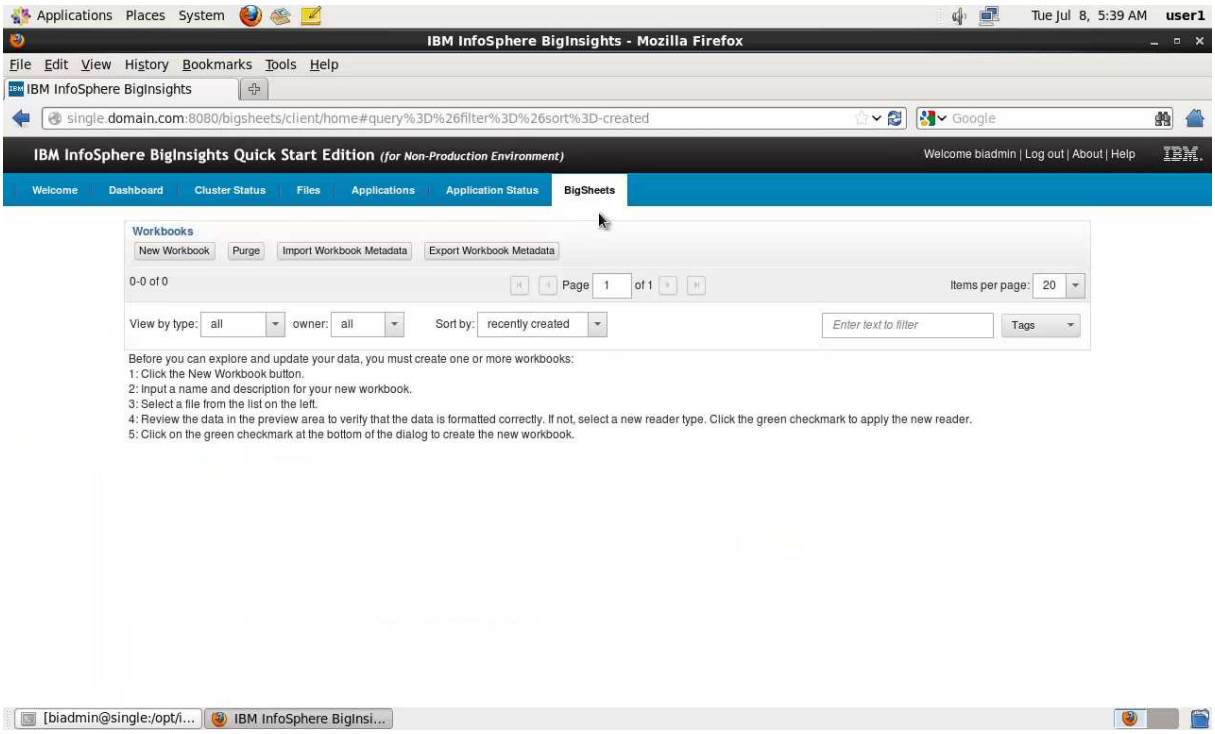


Abbildung 14: Anzeige des Tabs BigSheets

Dort sehen wir noch eine recht leere Arbeitsfläche, was daran liegt, dass wir noch kein sogenanntes Workbook angelegt haben. Generell werden uns bei der Arbeit mit BigSheets drei Begriffe bzw. Objekte begleiten:

- **Workbook**: Ein Workbook ist technisch gesehen eine Tabelle, die uns eine bestimmte Sicht auf Daten im HDFS oder in einer Datenbank zeigt. Ein Workbook legt dabei nie selber Daten an, sondern schaut nur aus einer bestimmte Perspektive auf die Daten. Ein Sonderfall des Workbooks ist das Master-Workbook, dessen Sichtweise nicht beeinflusst werden kann denn dieses bildet die dem Workbook zugrunde liegenden Daten immer genau so ab, wie sie im HDFS / der Datenbank vorliegen.
- **Sheet**: Auf Basis eines Workbooks können nun mehrere Sheets erstellt werden. Jedes Sheet hat dabei eines bestimmte Funktion, sei es z.B. ein Filter oder eine Gruppierung gemäß einer bestimmte Spalte eines Workbooks. Wir werden die Verwendung dieser Filter später noch genauer kennenlernen. Wenn ein Workbook mehrere Sheets beinhaltet, dann kann immer genau ein Sheet ausgewählt werden, dessen Daten die Daten des Workbooks repräsentieren. Ein Workbook stellt also immer die Sicht auf die Daten gemäß eines Sheets da.
- **Chart**: Daten eines Workbooks können über ein Chart visualisiert werden. Charts können etwa Diagramme wie Torten- oder Balkendiagramme sein oder auch neumodische Typen wie Bubble-Charts oder Landkarten.

Wir klicken nun auf *New Workbook*, geben Sie dem Workbook den Namen CustomersWB und wählen im neuen Fenster die zuvor hochgeladene Datei *customers.csv* aus dem Ordner */apps/customeranalysis/input* aus.

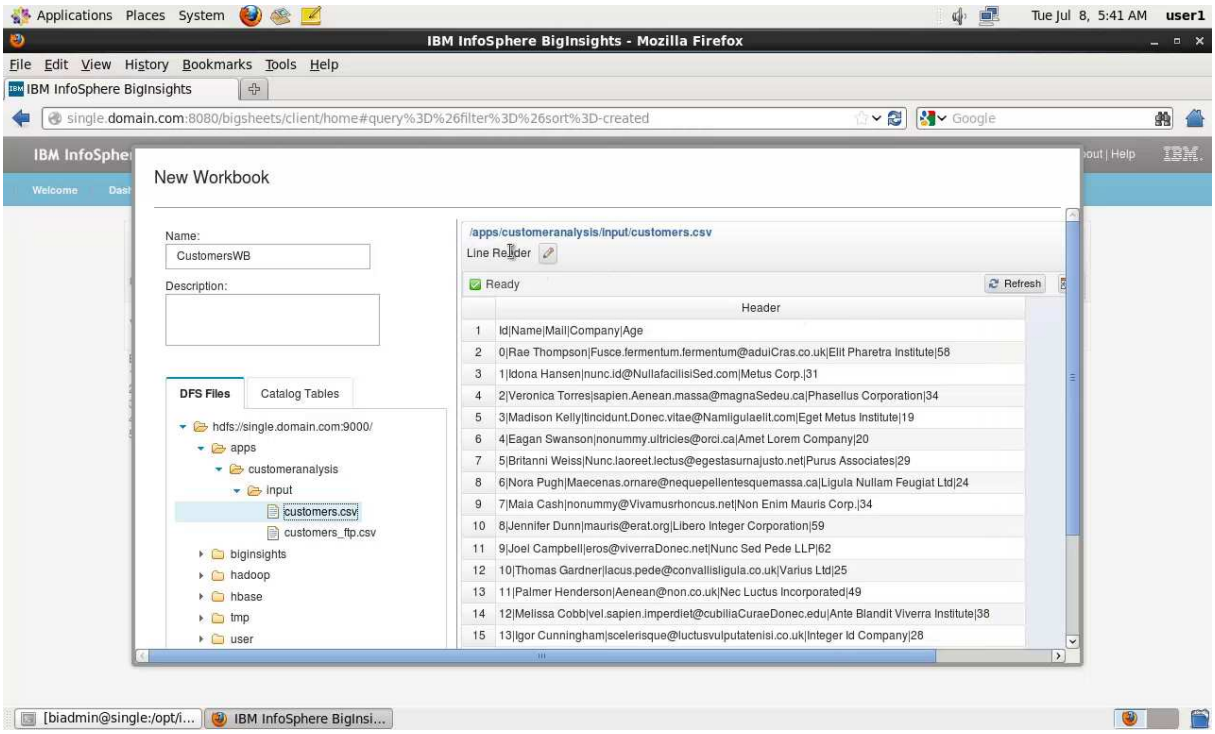


Abbildung 15: Anlegen eines neuen Workbooks

Im rechten Fenster sehen Sie, dass die Datei in einer Art Vorschau geladen wird. Da unser CSV noch falsch interpretiert wird, müssen wir *BigSheets* mitteilen, über welche Art von *Reader* es verwenden soll, um die Daten zu interpretieren.

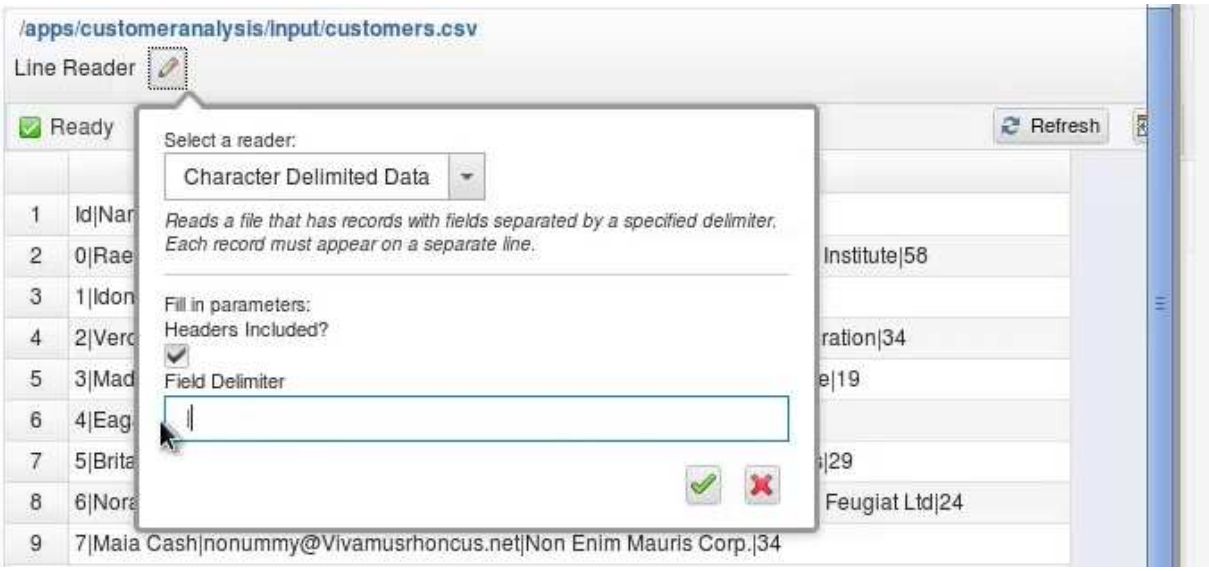
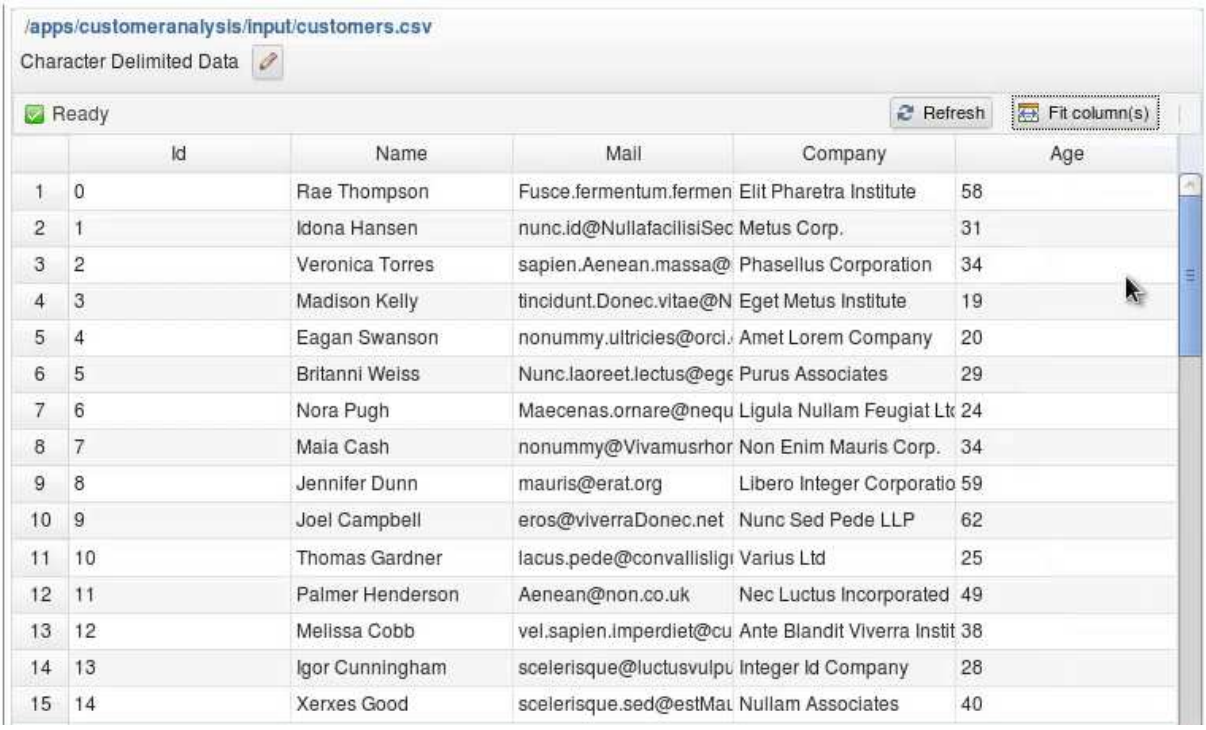


Abbildung 16: Auswahl des passenden Readers

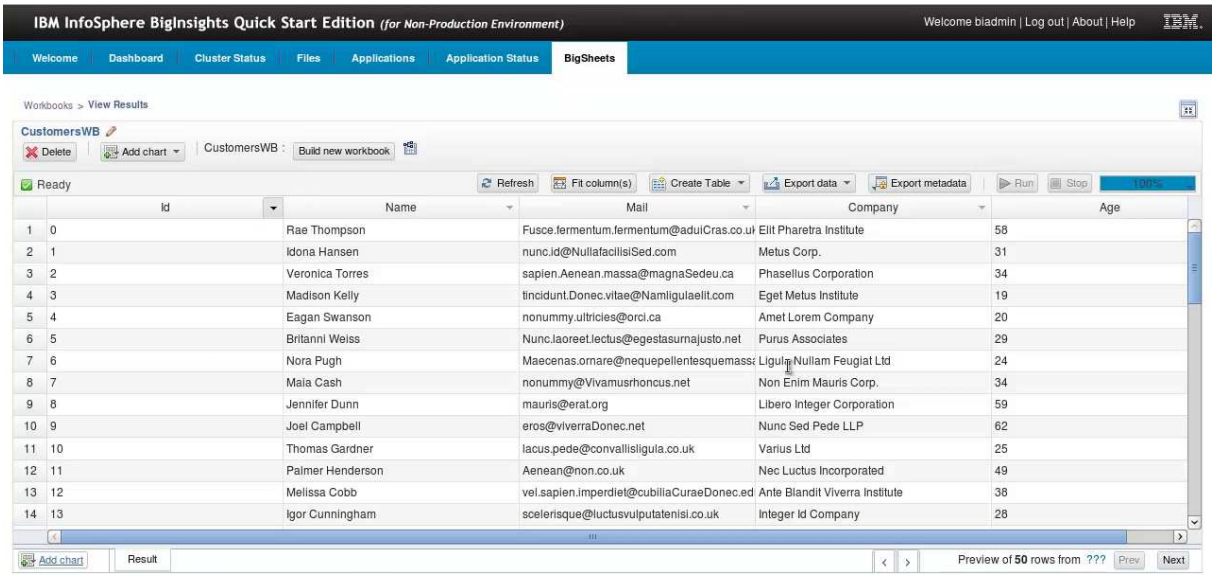
Wir verwenden den *Character Delimited Data Reader*, der in der Lage ist, CSV-ähnliche Dateien einzulesen, die nun nicht mit einem Komma, sondern mit einem senkrechten Strich | getrennt sind. Das Häkchen bei *Headers Included?* belassen wir, denn unsere Datei beinhaltet ja tatsächlich eine Kopfzeile. Klicken Sie abschließend das grüne Häkchen an, wird der Reader übernommen. Mit einem Klick auf Fit columns in der oberen, rechten Ecke, werden die Spaltenbreiten zusätzlich noch automatisch gesetzt und Sie sollten eine Ansicht ähnlich der folgenden erhalten.



	Id	Name	Mail	Company	Age
1	0	Rae Thompson	Fusce.fermentum.fermen	Elit Pharetra Institute	58
2	1	Idona Hansen	nunc.id@NullafacilisiSec	Metus Corp.	31
3	2	Veronica Torres	sapien.Aenean.massa@	Phasellus Corporation	34
4	3	Madison Kelly	tincidunt.Donec.vitae@N	Eget Metus Institute	19
5	4	Eagan Swanson	nonummy.ultricies@orci.	Amet Lorem Company	20
6	5	Britanni Weiss	Nunc.laoreet.lectus@ege	Purus Associates	29
7	6	Nora Pugh	Maecenas.ornare@nequ	Ligula Nullam Feugiat Lt	24
8	7	Maia Cash	nonummy@Vivamusrhor	Non Enim Mauris Corp.	34
9	8	Jennifer Dunn	mauris@erat.org	Libero Integer Corporatio	59
10	9	Joel Campbell	eros@viverraDonec.net	Nunc Sed Pede LLP	62
11	10	Thomas Gardner	lacus.pede@convallisligi	Varius Ltd	25
12	11	Palmer Henderson	Aenean@non.co.uk	Nec Luctus Incorporated	49
13	12	Melissa Cobb	vel.sapien.imperdiet@cu	Ante Blandit Viverra Instit	38
14	13	Igor Cunningham	scelerisque@luctusvulpu	Integer Id Company	28
15	14	Xerxes Good	scelerisque.sed@estMau	Nullam Associates	40

Abbildung 17: Formatierte Eingabedaten

Wie Sie sehen, wurden die Spaltenbeschriftungen ebenfalls übernommen. Scrollen Sie in dem Fenster ganz nach unten und schließen Sie somit die Erstellung des Workbooks ab. Es öffnet sich dann sogleich und präsentiert uns die eben importierten Daten erneut.



	Id	Name	Mail	Company	Age
1	0	Rae Thompson	Fusce.fermentum.fermentum@aduiCras.co.uk	Elit Pharetra Institute	58
2	1	Idona Hansen	nunc.id@NullafacilisiSed.com	Metus Corp.	31
3	2	Veronica Torres	sapien.Aenean.massa@magnaSedeu.ca	Phasellus Corporation	34
4	3	Madison Kelly	tincidunt.Donec.vitae@Namligulaelit.com	Eget Metus Institute	19
5	4	Eagan Swanson	nonummy.ultricies@orci.ca	Amet Lorem Company	20
6	5	Britanni Weiss	Nunc.laoreet.lectus@egestasurnajusto.net	Purus Associates	29
7	6	Nora Pugh	Maecenas.ornare@nequepellentesquemassa	Ligula Nullam Feugiat Ltd	24
8	7	Maia Cash	nonummy@Vivamusrhoncus.net	Non Enim Mauris Corp.	34
9	8	Jennifer Dunn	mauris@erat.org	Libero Integer Corporation	59
10	9	Joel Campbell	eros@viverraDonec.net	Nunc Sed Pede LLP	62
11	10	Thomas Gardner	lacus.pede@convallisligula.co.uk	Varius Ltd	25
12	11	Palmer Henderson	Aenean@non.co.uk	Nec Luctus Incorporated	49
13	12	Melissa Cobb	vel.sapien.imperdiet@cubiliaCuraeDonec.ed	Ante Blandit Viverra Institute	38
14	13	Igor Cunningham	scelerisque@luctusvulputatenisi.co.uk	Integer Id Company	28

Abbildung 18: Ansicht der Daten im Workbook

Ein Klick auf den Link *Workbooks* oben links vor dem *View Results* bringt uns zurück auf die Übersichtsseite. Dort sehen Sie nun unser CustomersWB. Das Tabellensymbol mit dem kleinen Schloss daneben, weist darauf hin, dass es sich um ein Master-Workbook handelt und das die sich darin befindlichen Daten bzw. die Sicht darauf, nicht verändern lassen.

Klicken Sie das Workbook erneut an, um es zu öffnen. Über den Button *Build New Workbook* rechts neben *CustomersWB* erzeugen wir nun ein neues Workbook. Dieses trägt zu Beginn den Namen

CustomersWB(1). Um diesen unschönen Namen durch einen passenden zu ersetzen, klicken wir auf den Bleistift oben links neben den Namen und geben in dem kleinen Fenster *YoungCustomers* ein. Der Name lässt schon vermuten, was wir im nächsten Schritt erreichen wollen. Wir wollen in dem neuen Workbook eine Sicht auf die Kundendaten erlangen, die alle jungen Kunden mit einem Alter unter 30 auflistet.

Klicken Sie dazu oben auf *Add sheets* und wählen Sie in der Drop-Down-Liste den Eintrag *Filter* aus.

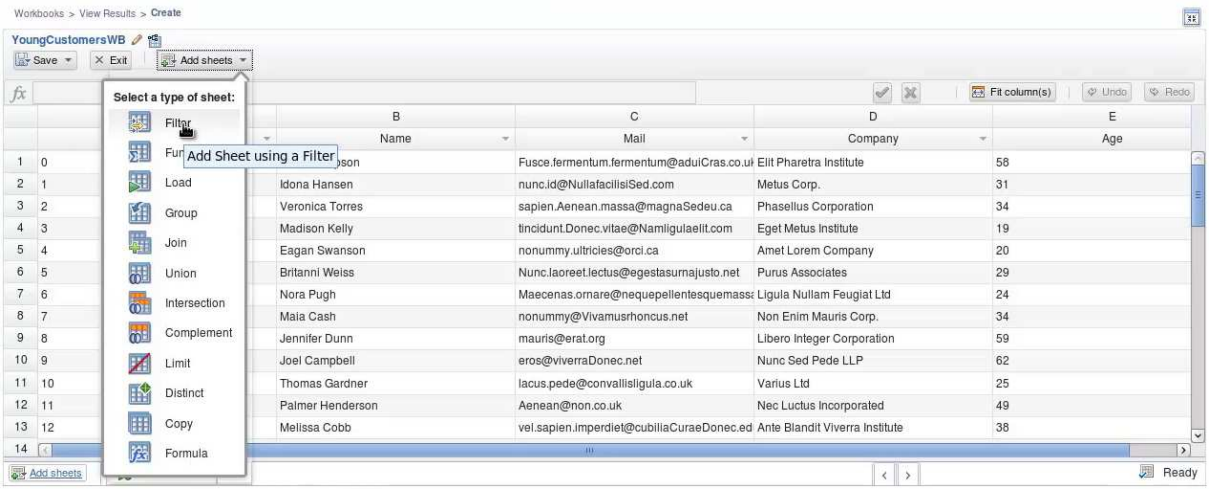


Abbildung 19: Anlegen eines Sheets zum Filtern von Daten

Im nächsten Fenster können Sie das Sheet nun benennen, ich habe mal den Namen *YoungCustomersSheet* gewählt. Darauf folgt eine Auswahl von Spalten, Vergleichsoperatoren und Vergleichswerten. In der ersten Drop-Down-Box legen wir die Spalte Age fest auf Basis derer unser Filter arbeiten soll. In der zweiten Auswahl wählen wir *smaller than* und in der dritten die Zahl 30. Die dritte Drop-Down-Box bietet uns schon einige Werte an, die es anhand der Daten ermittelt, die in den ersten Zeilen der Tabelle vorgekommen sind. Wenn der gewünschte Wert nicht dabei ist, können Sie diesen auch manuell eintragen.

Abbildung 20: Parametrisierung des Filters

Klicken Sie abschließend auf den kleinen grünen Pfeil unten und der Filter wird angewandt. Zurück im Workbook selektieren Sie bitte anschließend oben Links *Save & Exit*. Bestätigen Sie das Speichern mit *Save*. Wenn wir nun den Editierungsmodus verlassen, dann bekommen wir eine Meldung angezeigt, die uns auffordert, das Workbook über *Run* erstellen zu lassen. Der Hintergrund dabei ist der, dass

die Sicht in Form von einem Workflow, bestehend aus mehreren Map-Reduce-Jobs, durchgeführt werden muss. Während das geschieht, können Sie ganz normal weiterarbeiten. Vielleicht wirkt es ein wenig verwunderlich, dass dieser Schritt notwendig ist, schließlich haben wir ja bereits in der Tabelle Daten gesehen. Das ist zwar richtig, jedoch verwendet BigSheets nur eine Vorschau auf den gesamten Datensatz (etwa die ersten 50 Zeilen). Wenn man bedenkt, dass man in einer Big-Data-Plattform in der Regel mit Dateien in der Größenordnung von Gigabytes arbeitet, ist das auch sehr sinnvoll. So können Sie ihre Analysen konzipieren, ohne ständig auf die Jobs im Hintergrund zu warten, die die großen Datenmengen ständig aufbereiten müssen.

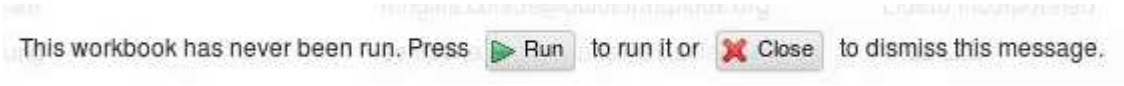


Abbildung 21: Aufbereiten aller Daten im Workbook über Run

Lange Rede, kurzer Sinn: Klicken Sie bitte auf *Run* und BigSheets erzeugt das Workbook und verwendet diesmal alle Daten unserer *customers.csv*. Wenn Sie stattdessen auf *Close* klicken, dann arbeiten Sie weiterhin mit der Datenvorschau.

Diesen Aufbereitungsschritt können Sie auch jederzeit über das Run in der oberen rechten Ecke des Fensters vornehmen. Daneben befindet sich ebenfalls eine Anzeige, die den Fortschritt der Datenverarbeitung wiedergibt.

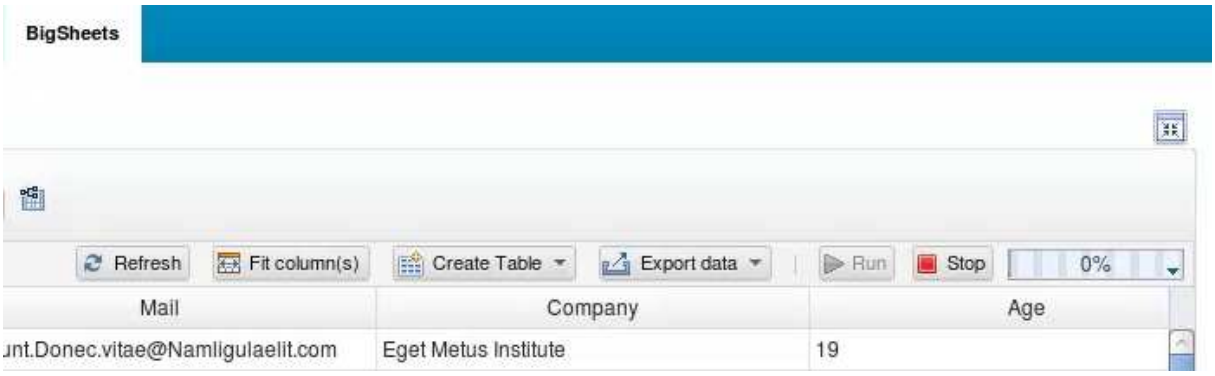


Abbildung 22: Fortschritt der Datenaufbereitung

Wenn Sie in einem Workbook mehrere Sheets verwenden, so können Sie über die Fußzeile festlegen, welches Sheet im Workbook repräsentiert werden soll. Klicken Sie dazu das entsprechende Sheet unten an und setzen Sie ein Häkchen bei *Set as Result sheet*.

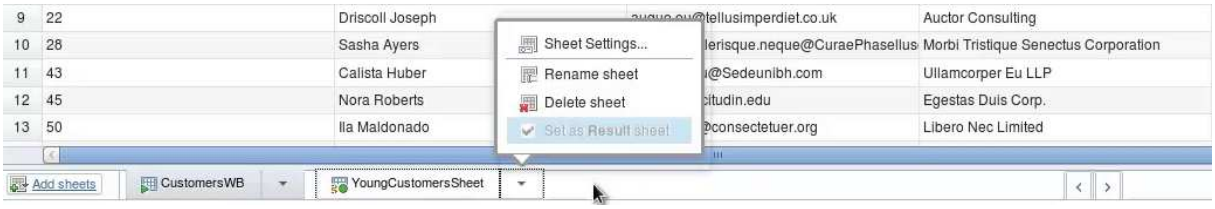


Abbildung 23: Setzen des Result-Sheets eines Workbooks

Verlassen Sie nun den Edit-Modus über den Button Exit im oberen Menü. Der Button *Add sheets* verschwindet und dafür taucht *Add chart* auf. Wir wollen nun im nächsten Schritt also ein Diagramm erzeugen. Es soll genauer gesagt ein Tortendiagramm werden, das uns anzeigt welche Altersstufen

unter unseren jungen Kunden am häufigsten vorkommen. So lautet dessen Aussage zum Beispiel: 3 Kunden sind 18 Jahre alt, 10 Kunden sind 19 Jahre alt, usw.

Klicken Sie nun auf *Add chart* und wählen Sie die Kategorie *Chart* und darunter *Pie*.

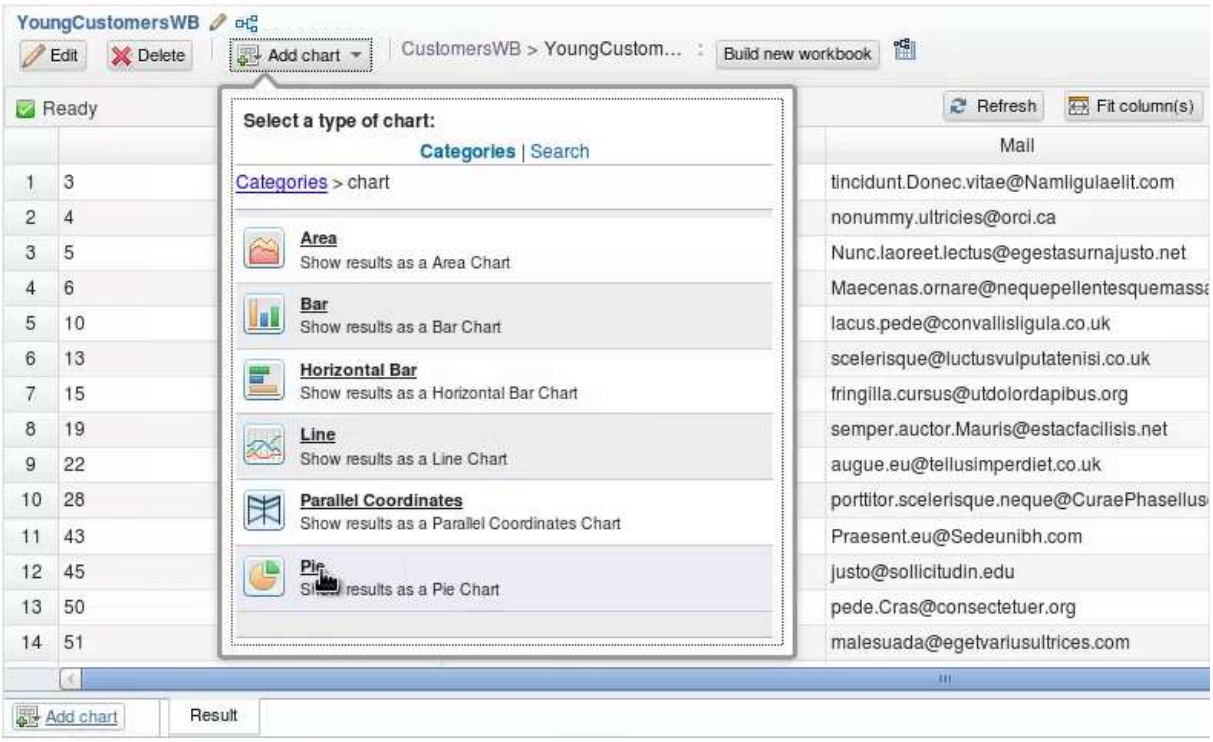


Abbildung 24: Einfügen eines Pie-Charts

Es öffnet sich ein Konfigurationsfenster, in dem Sie die Werte, die für das Zeichnen des Charts verwendet werden, festlegen können. Als Name sowie als Titel des Diagramms verwende ich die Bezeichnung *Age*, da wir die Altersstufen abbilden wollen. Der Wert, der in dem Tortendiagramm dargestellt werden soll, kommt aus der Spalte *Age*, die wir im Feld *Value* über die Drop-Down-Liste auswählen. Unter Count lassen wir die Bezeichnung *Count occurrences of X axis values*. Dadurch sagen wir, dass das Diagramm das Vorkommen jeder Altersstufe zählen und anzeigen soll. Die Werte sollen weiterhin aufsteigend geordnet werden und wir wollen sie auf 15 Einträge limitieren, um unser Diagramm nicht zu überladen. *Template* und *Style* belassen wir bei den Default-Werten.

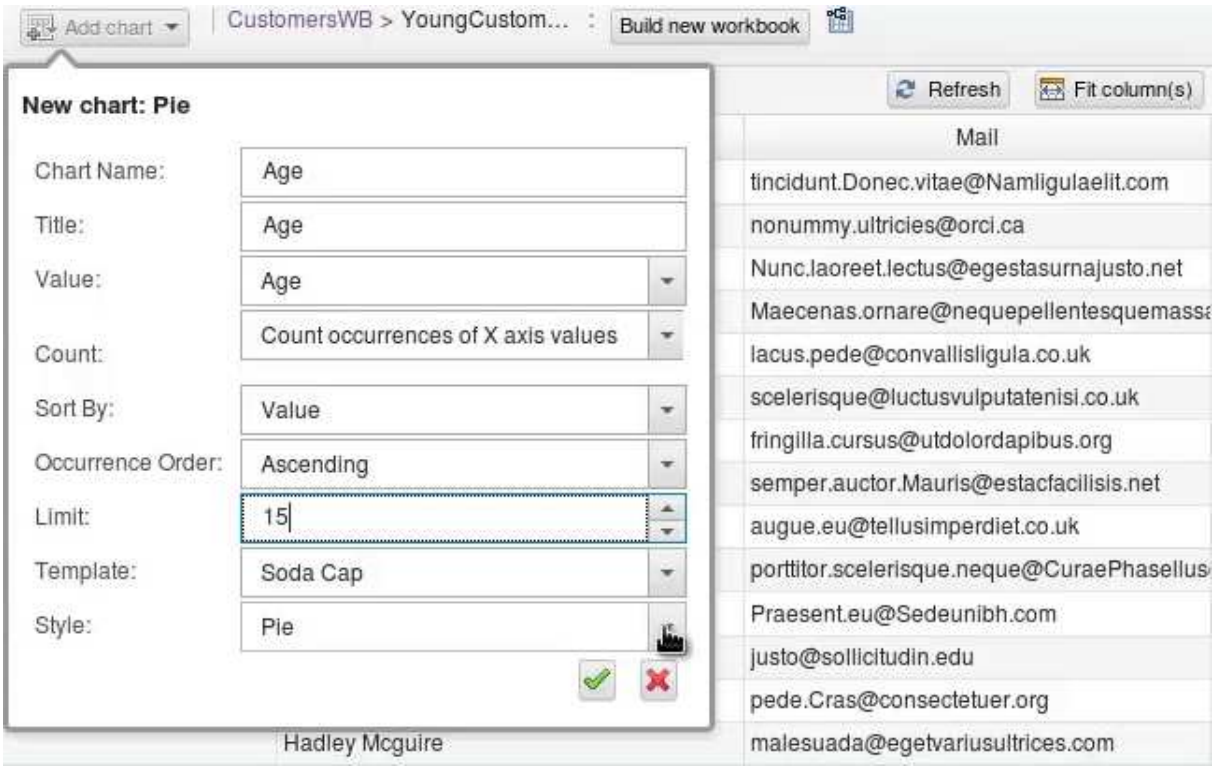


Abbildung 25: Konfigurieren des Charts

Bestätigen Sie die Konfiguration über den grünen Pfeil und das Diagramm wird sogleich geladen. Ebenso wie ein Sheet, muss es über *Run* noch erzeugt werden. Die Vorschau des Diagramms basiert wie auch bei einem Sheet aus einer Vorschau der darzustellenden Daten.

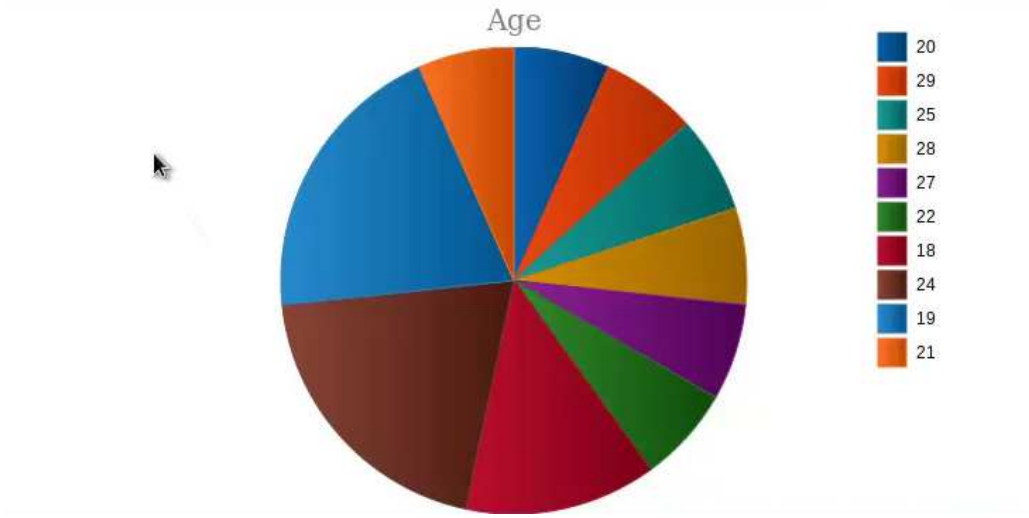


Abbildung 26: Ein Pie-Chart visualisiert Altersstufen im Diagramm

Es ist zu sehen, dass der größte Anteil der jungen Kunden 24 und 19 Jahre alt ist. Wenn Sie mit der Maus über das Diagramm fahren, dann werden Ihnen die einzelnen Kategorien numerisch neben der Maus angezeigt.

Klicken Sie nun oben Links erneut auf *Workbooks*, um in die Übersicht über alle Workbooks zu geladen. Klicken Sie dann neben *CustomersWB* auf das kleine Diagramm mit dem Tool-Tip *Workflow Diagram*.

Abbildung 27: Ansicht der Abhängigkeiten der Workbooks

Hier sehen Sie, welche Workbooks und Charts aufeinander aufbauen bzw. voneinander abhängig sind.

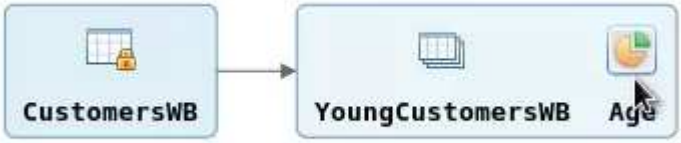


Abbildung 28: Abhängigkeiten der Workbooks

Das Workbook *YoungCustomersWB* ist also vom *CustomersWB* abhängig und beinhaltet zudem ein Chart *Age*. Über diese Ansicht behalten Sie immer die Übersicht über alle Workbooks und sehen z.B., ob Sie ein WB entfernen können oder nicht.

Schließen Sie die Ansicht nun über das *X* in der oberen rechten Ecke und öffnen Sie erneut das *CustomersWB*. Darin wollen wir nun ein neues Workbook anlegen, dass uns das Durchschnittsalter der Kunden angibt, die bei einer bestimmten Firma arbeiten. So könnte eine Aussage lauten: Unseren Kunden, die beim Firma XY arbeiten, sind im Durchschnitt 20 Jahre alt.

Erzeugen Sie also über den Button *Build new workbook* in dem oberen Menü ein neues Workbook und benennen Sie es über einen Klick auf den Bleistift neben dem Namen des neuen Workbooks (*CustomerWB(1))* um in *AverageAgePerCompany*. Wie bereits erklärt, müssen wir nun ein Sheet erstellen, dass es uns erlaubt, die Datenanalyse wie gewünscht durchzuführen. Da wir zuerst die Daten nach Arbeitgeber unserer Kunden gruppieren, müssen wir in der Liste unter *Add Sheet* den Eintrag *Group* auswählen.

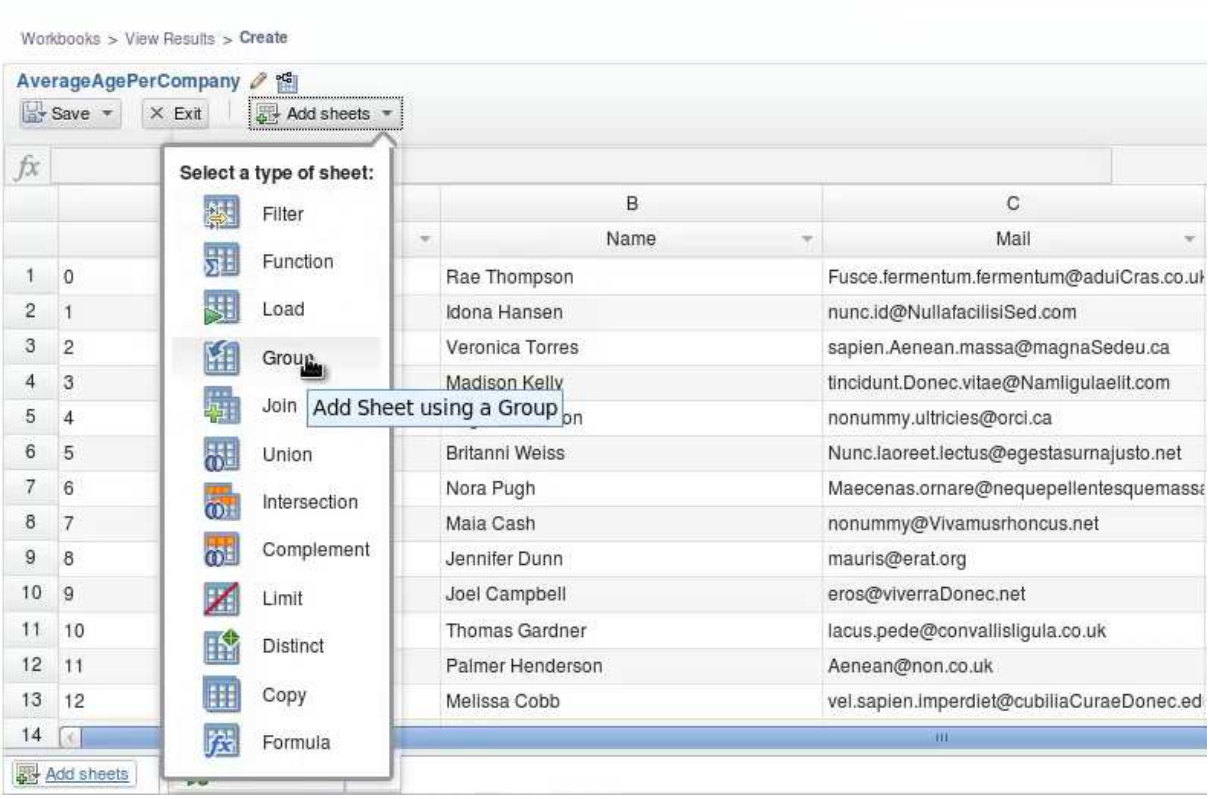


Abbildung 29: Erstellen eines Sheets auf Basis einer Gruppierung

Im folgenden Fenster spezifizieren wir die Gruppierung. Zuerst geben wir dem Sheet den Namen *AverageAgePerCompanySheet* und wählen dann in der Liste *Group by columns* den Eintrag *Company* aus. Klicken Sie danach auf das grüne Plus, um dem Sheet die Gruppierung hinzuzufügen.

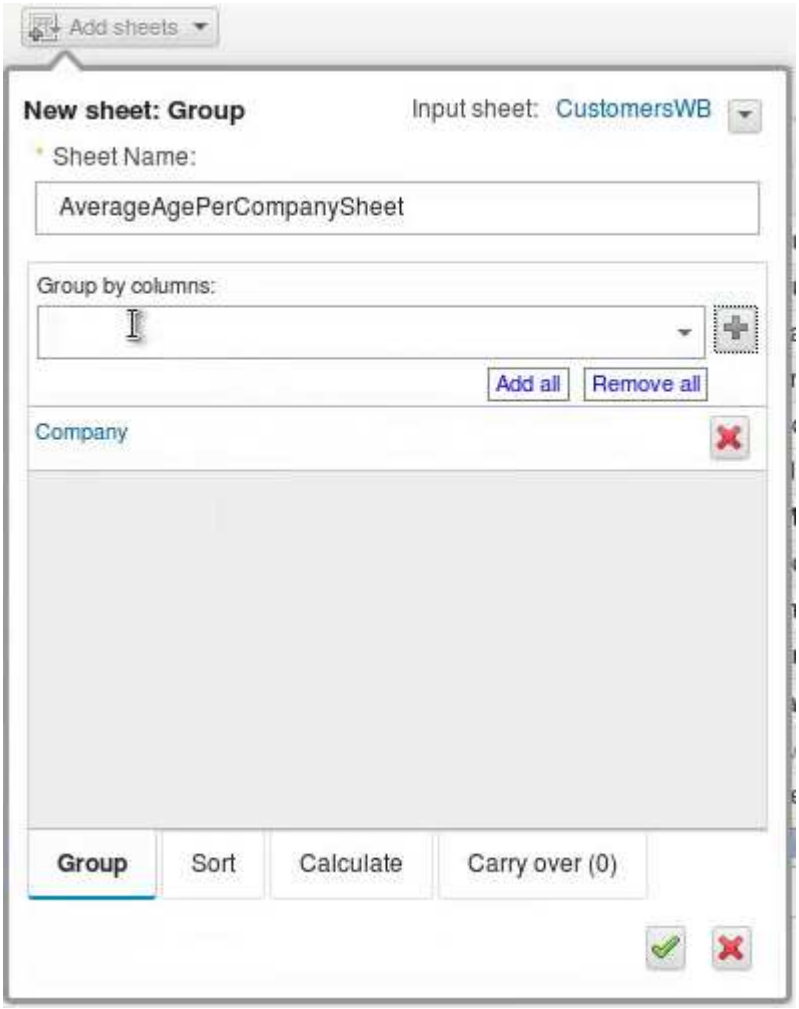


Abbildung 30: Hinzugefügte Gruppierung nach Company

Wechseln Sie dann in den Tab *Calculate* im unteren Bereich des Fensters. Dort definieren wir eine Spalte, die auf Basis einer Berechnung zum Sheet hinzugefügt wird. Tragen Sie also bei *Create columns based on groups* den Namen *Average_age* ein, um genau eine solche Spalte zu erzeugen. In dem nun erzeugten Eintrag wählen Sie aus der Liste neben dem Bezeichner *Average_age* = den Eintrag *AVG* aus. Alternativ könnten Sie auch z.B. das Minimum oder Maximum einer Spalte, oder auch deren Summe berechnen. Neben dem Feld *Column* wird nun eine Drop-Down-Liste aktiviert, für die die Berechnung eines Durchschnittswerts möglich ist. Dort finden Sie die Spalten *Id* und *Age* und wählen entsprechend unseres Szenarios *Age* aus.

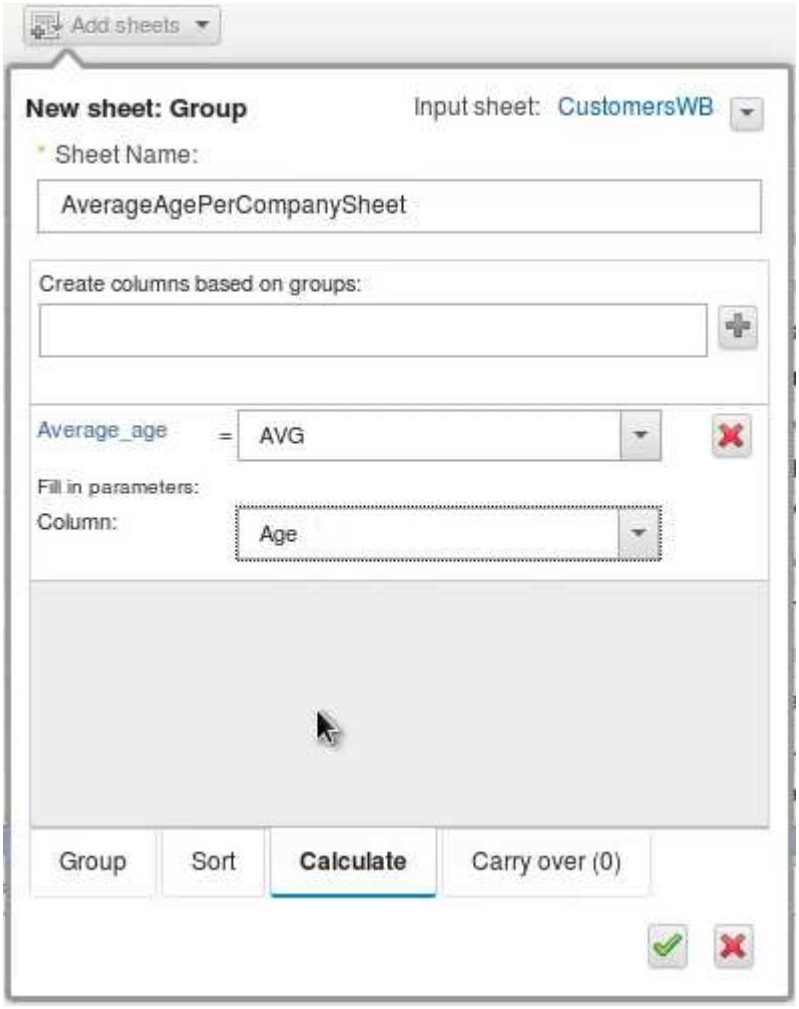


Abbildung 31: Berechnen der neuen Spalte Average_age

Klicken Sie zum Abschluss auf das grüne Häkchen, um das neue Sheet anzulegen. Klicken Sie dann oben auf *Save & Exit* und bestätigen Sie mit *Save*. Lassen Sie das Sheet dann entsprechend über *Run* generieren. Nach Abschluss sollte das Workbook wie in der folgenden Abbildung aussehen.

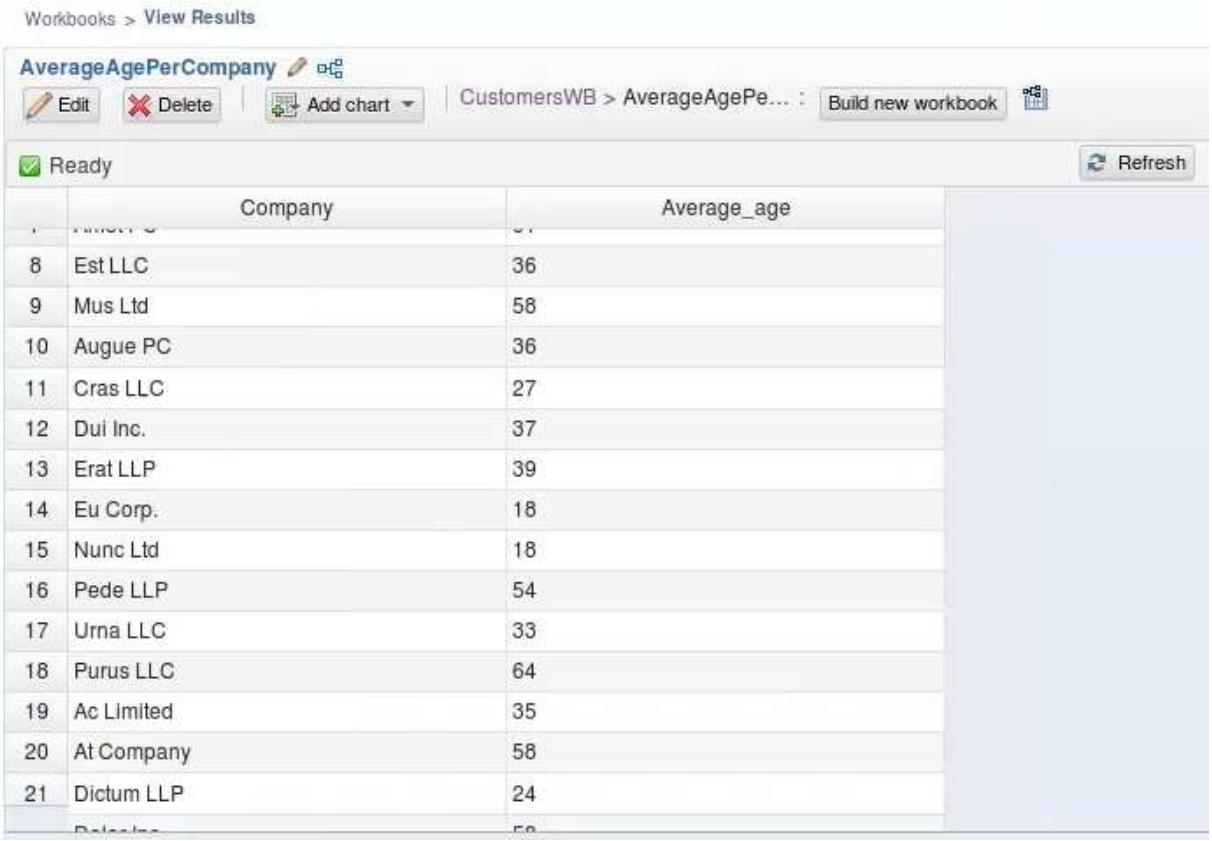


Abbildung 32: Durschnittsalter der Kunden nach Arbeitgeber

Es ist also zu sehen, dass unseren Kunden, die bei *Est LLC* arbeiten, im Durchschnitt 36 Jahre alt sind und diejenigen bei *Purus LLC* 64 Jahre. Diese Erkenntnis verlangt nach einer erneuten Visualisierung. Klicken Sie also oben auf *Add chart* und wählen Sie aus der Kategorie *Charts* den Typ *Line*.

Nennen Sie es beispielweise ComapnyAvgAge und geben Sie dem Diagramm einen beliebigen Titel. Auf der X-Achse sollen nun die verschiedenen Firmen aufgetragen werden, daher beschriften wir diese auch mit *Company*. Auf der Y-Achse soll unsere berechnete Spalte *Average_age* zusammen mit der Beschrifung *Average age* gezeigt werden. Die Einträge sollen aufsteigend geordnet sein und eine Anzahl von 20 nicht überschreiten, um das Diagramm nicht zu unübersichtlich werden zu lassen. *Template* und *Style* belassen wir bei den Ursprungswerten.

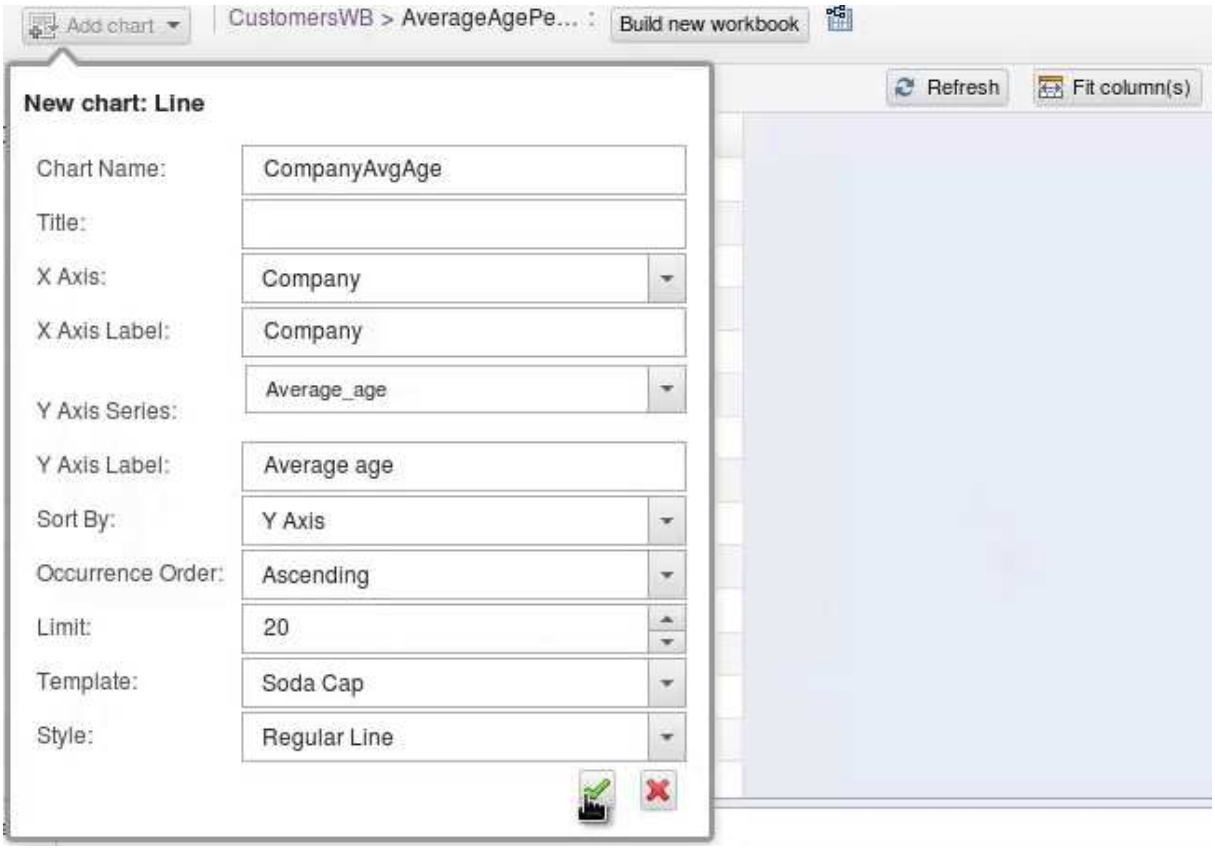


Abbildung 33: Definition eines Line-Charts

Klicken Sie zum Abschluss der Erstellen auf den grünen Pfeil und anschließend auf *Run*. Während das Diagramm nun die nötigen Daten zusammenstellt, sehen Sie in der Vorschau, wie es nach der Fertigstellung aussehen könnte.

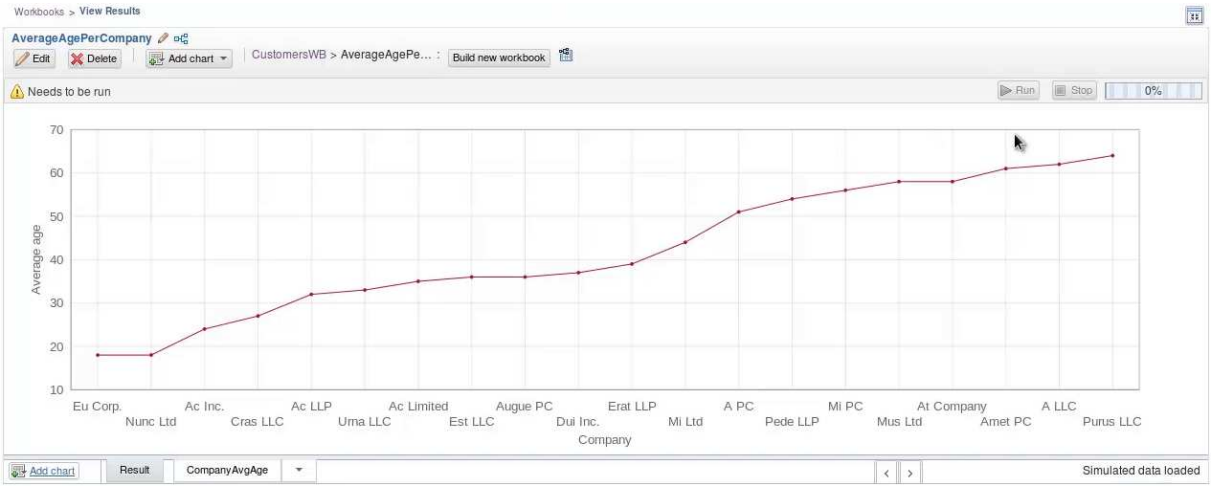


Abbildung 34: Vorschau auf die fertige Visualisierung des Durchschnittsalters der Kunden pro Arbeitgeber

Nun haben wir gesehen, wie wir Daten in BigSheets laden, analysieren und visualisieren. Versuchen Sie doch einmal mit den anderen Funktionen der Sheets zu arbeiten und probieren Sie ruhig noch einige weitere Diagrammtypen aus. Sie werden sehen, dass BigSheets ein mächtiges Werkzeug ist, um Big-Data ansehnlich und einfach aufzubereiten und auszuwerten.