Building OLAP on Hadoop

Intellicus Enterprise Reporting and BI Platform

Enterprise Reporting and Business Insights Platform

©Intellicus Technologies

info@intellicus.com
www.intellicus.com

Copyright © **2013** Intellicus Technologies

This document and its content is copyrighted material of Intellicus Technologies. The content may not be copied or derived from, through any means, in parts or in whole, without a prior written permission from Intellicus Technologies. All other product names are believed to be registered trademarks of the respective companies.

Dated: - May 2013.

Acknowledgements

Intellicus acknowledges using of third-party libraries to extend support to the functionalities that they provide.

For details, visit: http://www.intellicus.com/acknowledgements.htm

Contents

Working with OLAP on Hadoop

Intellicus takes its innovative multi-dimensional pre-aggregation based analytical solution to the Hadoop platform. Intellicus does the processing where the data resides; rather than bringing in data and then processing it. Using the new industry coveted capacity of Hadoop architecture to process peta bytes of data by parallel processing, Intellicus generates a series of map reduced jobs to pre-aggregate large amount of data, which already exists on Hadoop, without the need of bringing in data to Intellicus memory.

Intellicus also unleashes the power of NOSQL data stores to store and provide sub-second response to slicing and dicing from big data.

Supported Hadoop Environments

Following Hadoop environments are supported by Intellicus OLAP on Hadoop:

1. Apache hadoop 1.0.3 and above
2. Hortonworks 1.1.2 and above
3. MapR-M5

Installation of Intellicus on Hadoop

Prerequisites

The system requirements for Intellicus should include:

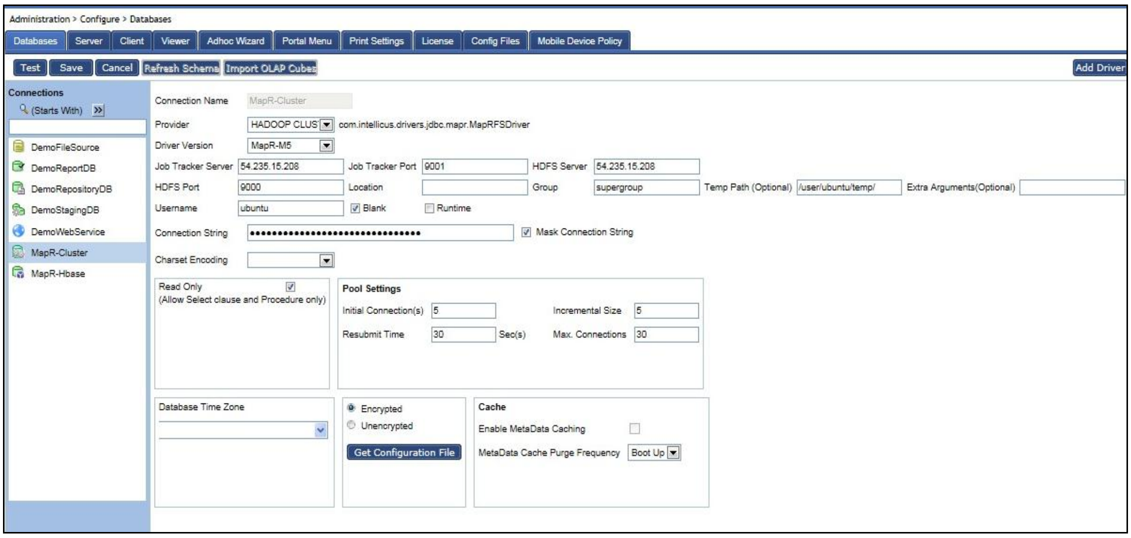
1. A linux machine, with any of the linux flavours (RH, Ubuntu, CentOS).
2. This machine should be on the same network on which the Hadoop cluster is running.
3. There should not be any firewalls or NAT between this machine and the Hadoop cluster machines.
4. The user with which the Intellicus server will be started shall have appropriate access rights on the Hadoop cluster.

Steps to Data Analytics on Hadoop

Creating a connection from Intellicus to Hadoop-Cluster

Below are the steps to create a connection to Hadoop cluster:

1. Go to **Administration > Configure > Databases**.
2. Click the **Add** button to create a new connection and provide the required information.



Connection Name: MapR-Cluster
Provider: HADOOP CLUS com.intellicus.drivers.jdbc.mapr.MapRFSDriver
Driver Version: MapR-M5
Job Tracker Server: 54.235.15.208
Job Tracker Port: 9001
HDFS Server: 54.235.15.208
HDFS Port: 9000
Location:
Group: supergroup
Temp Path (Optional): /user/ubuntu/temp/
Extra Arguments(Optional):
Username: ubuntu ☑ Blank ☐ Runtime
Connection String: •••••••••••••••••••••••••••••••• ☑ Mask Connection String
Charset Encoding:
Read Only ☑ (Allow Select clause and Procedure only)

Pool Settings
Initial Connection(s): 5
Incremental Size: 5
Resubmit Time: 30 Sec(s)
Max. Connections: 30

Database Time Zone:
● Encrypted
○ Unencrypted

Cache
Enable MetaData Caching ☐
MetaData Cache Purge Frequency: Boot Up

Figure 1: Creating Hadoop-Cluster Connection

The following table lists screen properties in order to create a connection to Hadoop-cluster:

Property	**Values**	**Comments**
Connection Name	Type Yourself	Name of the database connection
Provider	HADOOP CLUSTER	Data provider used for the connection
Driver Version	Select from the list	Version of Hadoop Cluster
Job Tracker Server	Type Yourself	IP address of Job Tracker Server
Job Tracker Port	Type Yourself	Port of Job Tracker Server
HDFS Server	Type Yourself	IP address of HDFS named node
HDFS Port	Type Yourself	Named node Port
Location	Type Yourself (Optional)	Default path for the connection on this HDFS Blank = location is root
Group	Type Yourself	User group on HDFS
User name	Type Yourself	User name on HDFS
Connection String	System Generated	Connection String to connect to HDFS
Charset Encoding	Select from the list	Select UTF-8 if the database is created or started with UTF-8 encoding. Otherwise, leave it blank
Read Only	Check/Uncheck the box	Check this checkbox to make sure only SQLs having read operations are executed on this connection

Pool Settings	Initial Connection(s)	Type Yourself	Specify the number of connections that should be opened initially. Default: 5
	Incremental Size	Type Yourself	Specify the number of connections to open when the all open connections are consumed. Default: 5
	Resubmit Time	Type Yourself	Specify the waiting-time in seconds before generating re-submitting request. Default: 30 seconds
	Max Connections	Type Yourself	Specify the maximum number of connections that can be opened to the selected database at a time. Default: 30
Database Time Zone		Select from the list	Select time zone to receive output of date / time fields as per time zone in which the data was entered in database
Get Configuration File		Click the Button	You can get the configuration file of Report Server in either Encrypted or Unencrypted format.
Cache	Enable Metadata Caching	Check/Uncheck the box	Check = The list of table names with column names (or other data source objects) from this connection will be pulled and stored locally for populating in SQL Editor or respective Query designer screens.
	MetaData Cache Purge Frequency	Select from the list: NEVER, BOOTUP	MetaData Cache Purge Frequency defines the time when metadata (table names, field names, etc.) cached for the selected connection should be deleted and refreshed: *NEVER* = application will never delete and refresh the metadata. BOOTUP= If this is set to *Boot Up*, every time server is booted, metadata for this data connection will be deleted and refreshed.

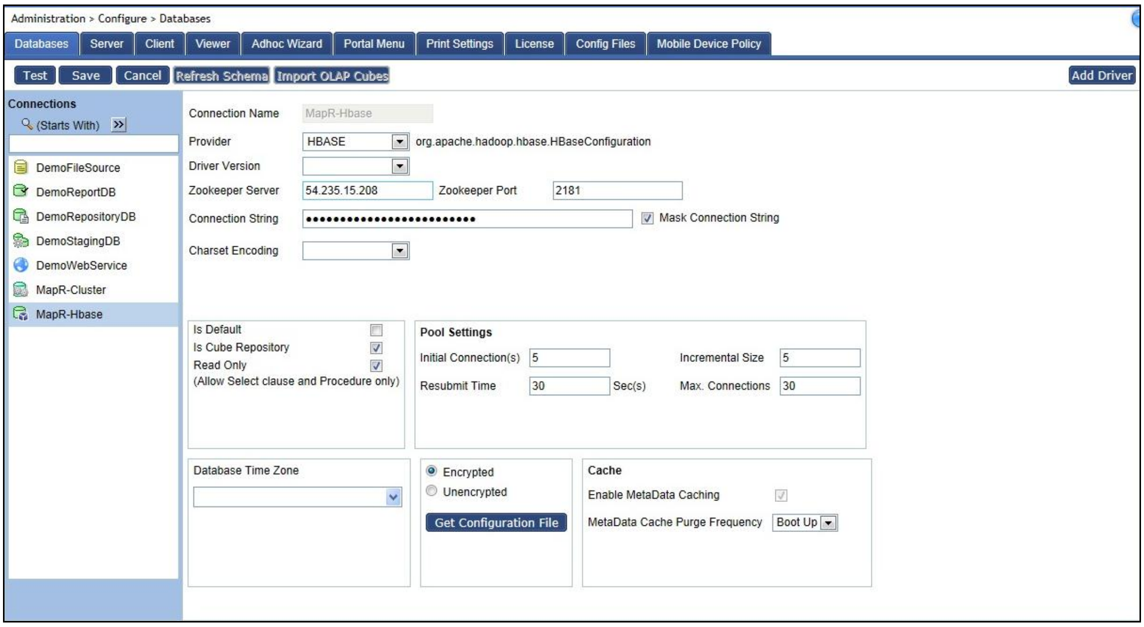
Action Buttons

Add	To start configuring a new connection
Modify	To modify selected connection
Delete	To delete selected connection
Refresh Schema	To manually refresh schema of the selected connection
Import OLAP Cubes	For OLAP type connections, open the dialog to import third party cubes

Creating an Hbase connection

Next you need to create an Hbase Connection:

1. Go to **Administration > Configure > Databases**.
2. Click the **Add** button to create a new connection and provide the required information.



Connection Name: MapR-Hbase
Provider: HBASE org.apache.hadoop.hbase.HBaseConfiguration
Driver Version:
Zookeeper Server: 54.235.15.208 Zookeeper Port: 2181
Connection String: •••••••••••••••••••••••• ☑ Mask Connection String
Charset Encoding:

Is Default □
Is Cube Repository ☑
Read Only ☑
(Allow Select clause and Procedure only)

Pool Settings
Initial Connection(s): 5 Incremental Size: 5
Resubmit Time: 30 Sec(s) Max. Connections: 30

Database Time Zone:

● Encrypted
○ Unencrypted

Cache
Enable MetaData Caching ☑
MetaData Cache Purge Frequency: Boot Up

Figure 2: Creating HBase Connection

The following table lists screen properties in order to create an HBase connection:

Property	**Values**	**Comments**
Connection Name	Type Yourself	Name of the database connection
Provider	HBASE	Data provider used for the connection
Driver Version	Select from the list	Version of HBase Server
Zookeeper Server	Type Yourself	IP address of Zookeeper Server
Zookeeper Port	Type Yourself	Port of Zookeeper Server
Connection String	System Generated	Connection String to connect to HBase
Charset Encoding	Select from the list	Select UTF-8 if the database is created or started with UTF-8 encoding. Otherwise, leave it blank
Is Default	Check/uncheck the box	Check this checkbox to use this connection as the application default data connection to fetch report data
Is Cube Repository	Check/uncheck the box	Check = Use this HBase connection as the data store to

			store the cubes built on Hadoop. This HBase database may need significant disk space depending on size of cubes.
Read Only		Check/Uncheck the box	Check = Report Server to make sure only SQLs having read operations are executed on this connection
Pool Settings	Initial Connection(s)	Type Yourself	Specify the number of connections that should be opened initially. Default: 5
	Incremental Size	Type Yourself	Specify the number of connections to open when the all open connections are consumed. Default: 5
	Resubmit Time	Type Yourself	Specify the waiting-time in seconds before generating re-submitting request. Default: 30 seconds
	Max Connections	Type Yourself	Specify the maximum number of connections that can be opened to the selected database at a time. Default: 30
Database Time Zone		Select from the list	Select time zone to receive output of date / time fields as per time zone in which the data was entered in database
Get Configuration File		Click the button	You can get the configuration file of Report Server in either Encrypted or Unencrypted format.
Cache	Enable Metadata Caching	Check/Uncheck the box	You can enable metadata caching
	MetaData Cache Purge Frequency	Select from the list	MetaData Cache Purge Frequency defines the time when metadata (table names, field names, etc.) cached for the selected connection should be deleted and refreshed: • If this is set to *Never*, application will never delete and refresh the metadata. • If this is set to *Boot Up*, every time server is booted, metadata for this data connection will be deleted and refreshed.

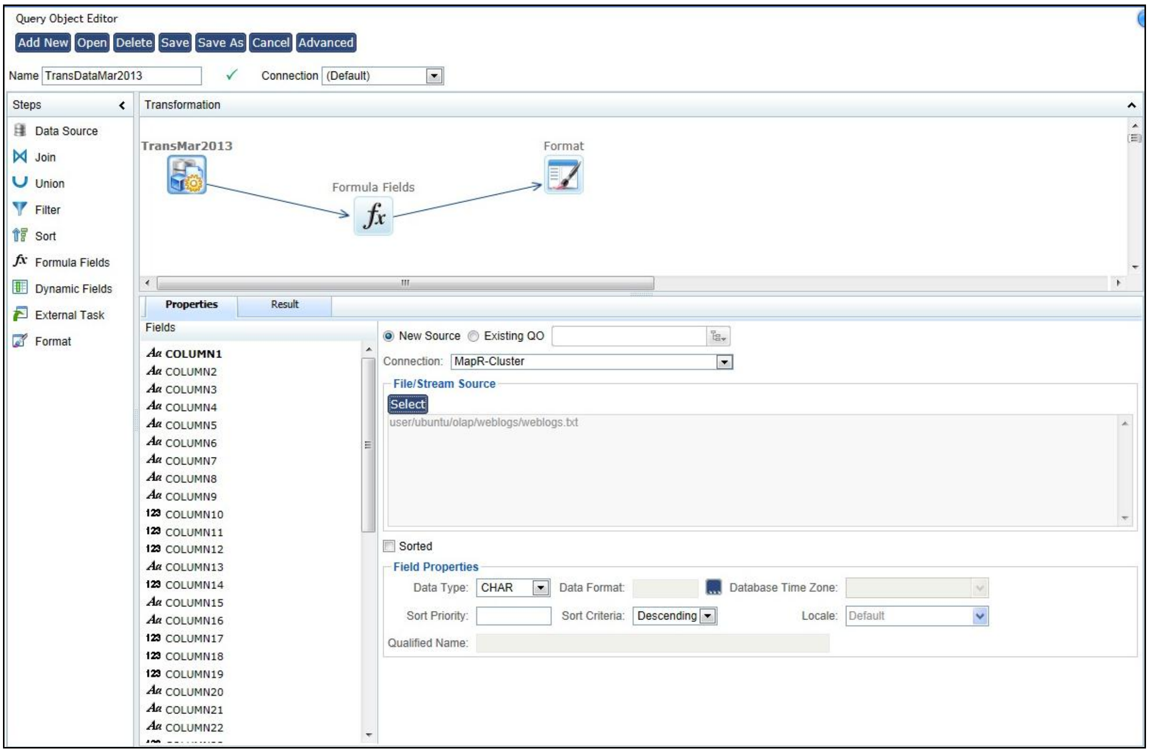
Action Buttons

Add	To start configuring a new connection
Modify	To modify selected connection
Delete	To delete selected connection
Refresh Schema	To manually refresh schema of the selected connection
Import OLAP Cubes	For OLAP type connections, open the dialog to import cubes

Creating a Query Object from file on HDFS

You will need to follow the below steps to create a query object:

1. Go to **Repository > Report Objects > Query**.
2. Click the **Add New** button to create a query object.



Name: TransDataMar2013 ✓ Connection: (Default)

● New Source ○ Existing QO

Connection: MapR-Cluster

user/ubuntu/olap/weblogs/weblogs.txt

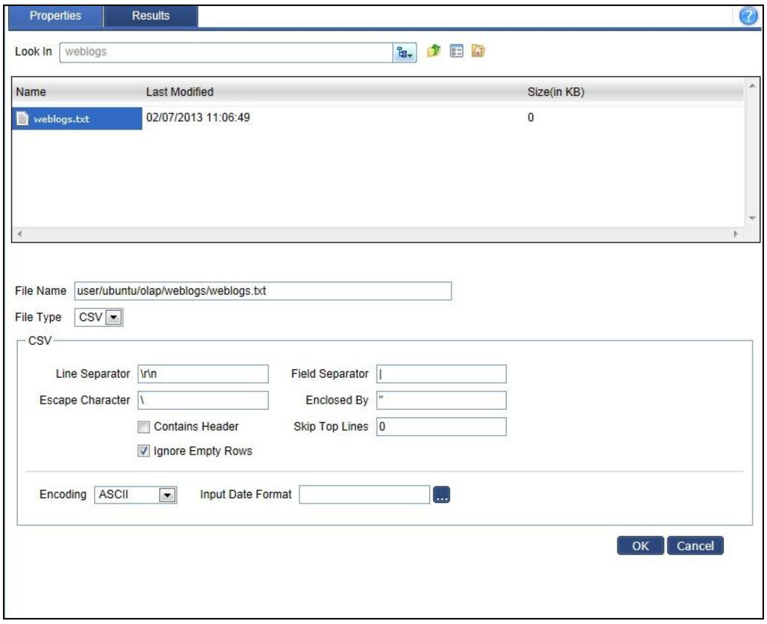
□ Sorted

Data Type: CHAR Data Format: Database Time Zone:

Sort Priority: Sort Criteria: Descending Locale: Default

Qualified Name:

Figure 3-a: Creating Query Object



Look In: weblogs

Name	Last Modified	Size(in KB)
weblogs.txt	02/07/2013 11:06:49	0

File Name: user/ubuntu/olap/weblogs/weblogs.txt

File Type: CSV

CSV

Line Separator: \r\n Field Separator: |

Escape Character: \ Enclosed By: "

□ Contains Header Skip Top Lines: 0

☑ Ignore Empty Rows

Encoding: ASCII Input Date Format:

Figure 4-b: Creating Query Object

The following steps help to create a Query Object:

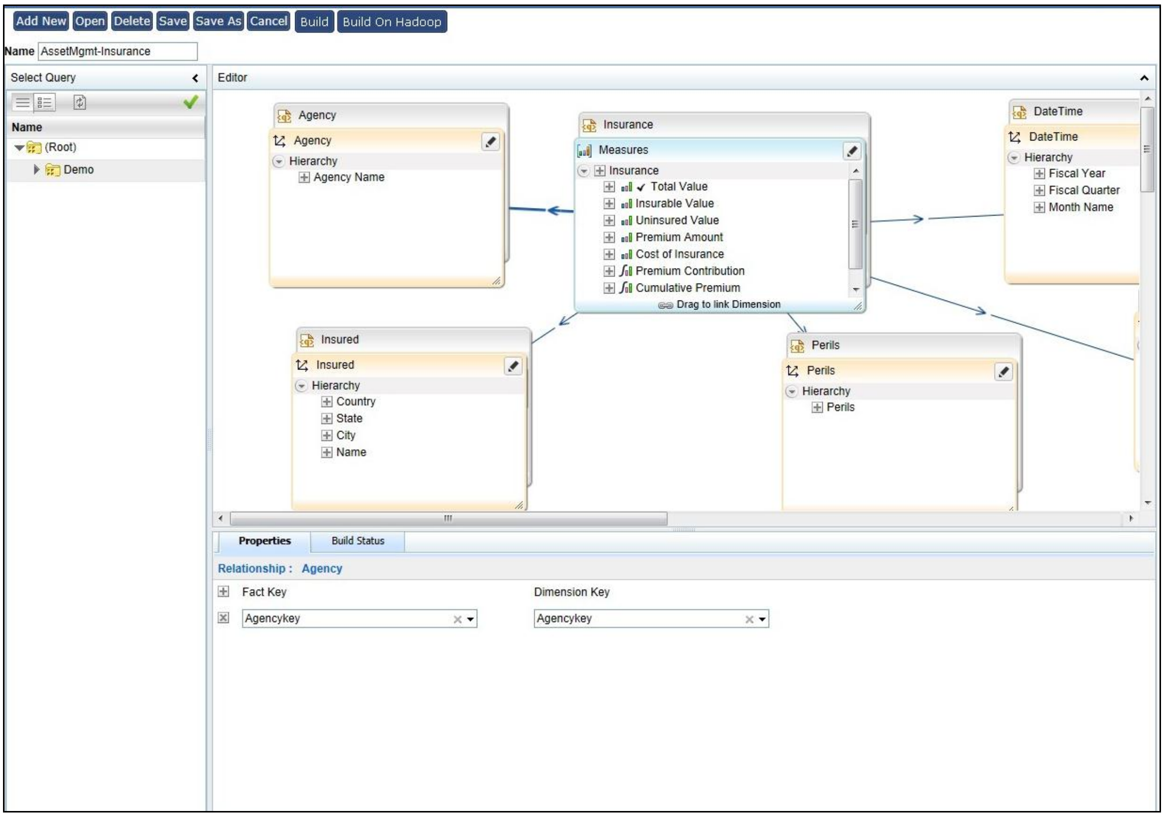
1. Provide a Query Object name in the **Name** field.
2. Set object level connection as Default.

3. Drag the Data Source from Steps pane and drop onto Transformation pane.
4. Set the step level connection as Hadoop-Cluster connection like MapR-cluster.
5. Click the **Select** button.
6. This would open up a new window (Figure 3-b) showing the HDFS file system of MapR-Cluster connection.
7. Select the desired file (either CSV or SEQ or XML) from the Look in directory.
8. Set the file settings as per the selected file type.
9. Click **OK** to view the selected file columns under Data Source section.
10. To perform certain transformations on your selected data like format, filter, formula and so on, please refer to "WorkingWithQueryObjects.pdf" from section **Simple Steps** onwards.

Creating a Cube Object using HDFS file Query Object

To create a cube object, you will need to follow the steps given below:

1. Go to **Repository > Report Objects > Cube**.
2. To create a Cube Object, please follow the steps as detailed under "WorkingWithOLAPCubes.pdf" from section **Designing Cube Objects** onwards.
3. If all the Query Objects are designed on the same connection of Hadoop, then upon re-opening the Cube, '**Build on Hadoop**' option will be enabled. '**Build on Hadoop**' builds on Hadoop enabled system when query objects are connected to Hadoop cluster, unlike '**Build**' that builds on H2 system database.



Name AssetMgmt-Insurance

Relationship : Agency

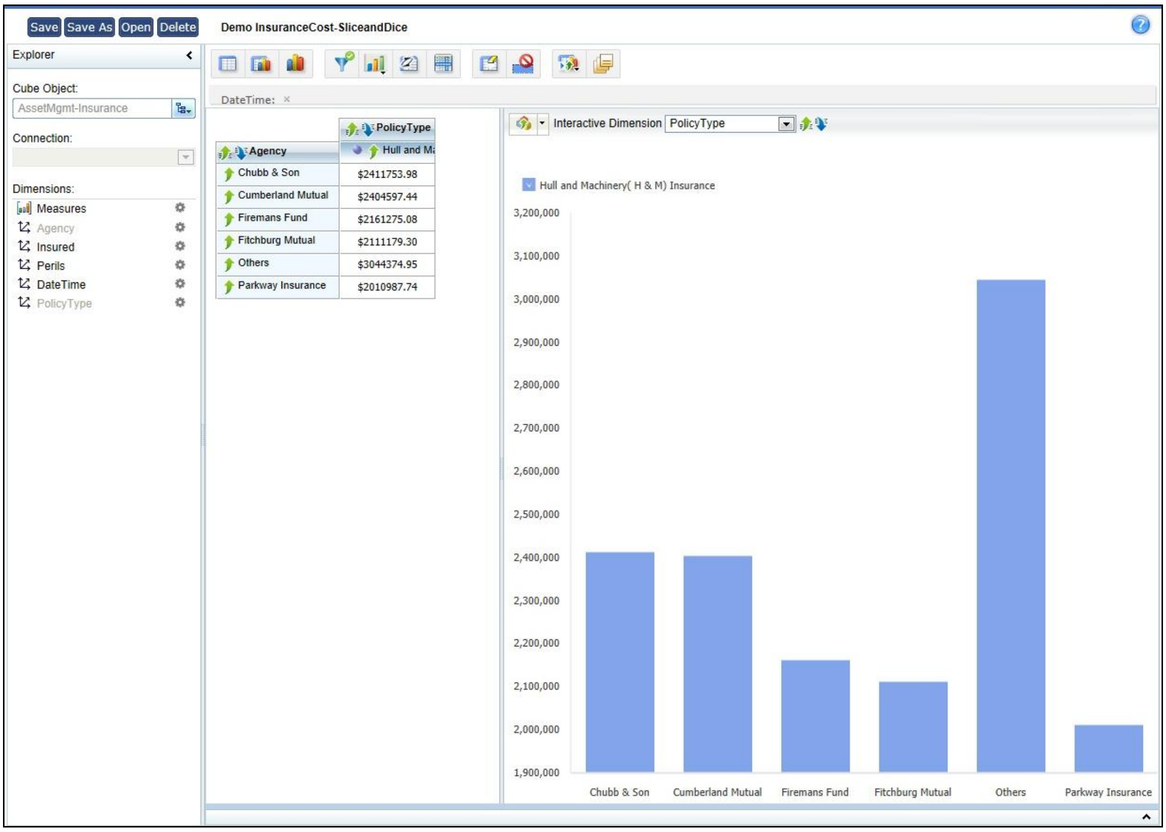
Fact Key	Dimension Key
Agencykey	Agencykey

Figure 4: 'Build on Hadoop' option on the Cube

Creating Visualizations

1. Login to the Intellicus Portal.
2. Go to **Navigation>Analytics> OLAP Viewer**.
3. Select Cube Object from the drop down list.
4. Select "**Measures**" from the available list.
5. Also, select "Dimensions" from the available list.

Based on your selection, you are now able to view and further analyse your data in either chart or grid or combination of chart/grid forms.



Demo InsuranceCost-SliceandDice

Cube Object: AssetMgmt-Insurance

Connection:

Dimensions:
- Measures
- Agency
- Insured
- Perils
- DateTime
- PolicyType

DateTime: ×

	PolicyType
Agency	Hull and Ma
Chubb & Son	$2411753.98
Cumberland Mutual	$2404597.44
Firemans Fund	$2161275.08
Fitchburg Mutual	$2111179.30
Others	$3044374.95
Parkway Insurance	$2010987.74

Interactive Dimension: PolicyType

Figure 5: Creating Visualizations