The Inside Scoop on Hadoop

Orion Gebremedhin

National Solutions Director – BI & Big Data , Neudesic LLC.
VTSP – Microsoft Corp.

Orion.Gebremedhin@Neudesic.COM
B-orgebr@Microsoft.com
@OrionGM

NEUDESIC

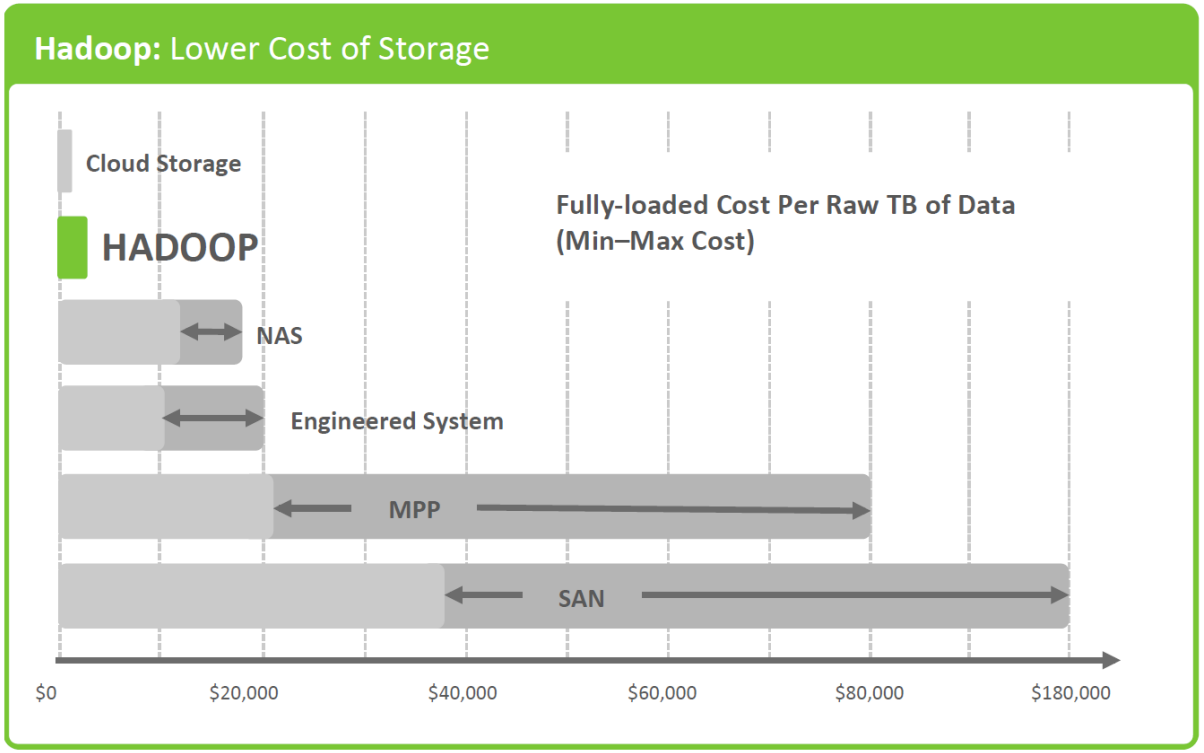
The Inside Scoop on Hadoop

Topics Covered

- Understanding Hadoop
- Big Data Solution Deployment Models
- Architecting the Modern Data Warehouse
- Summary

Understanding Hadoop

Big Data = Hadoop?



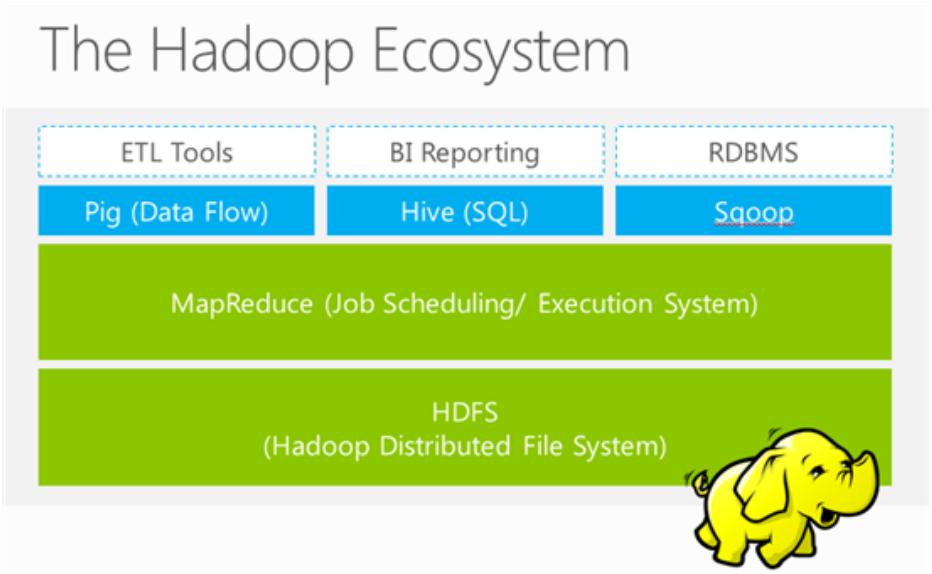
* A Modern Data Architecture with Apache Hadoop, Hortonworks Inc. 2014

The Fundamentals of Hadoop

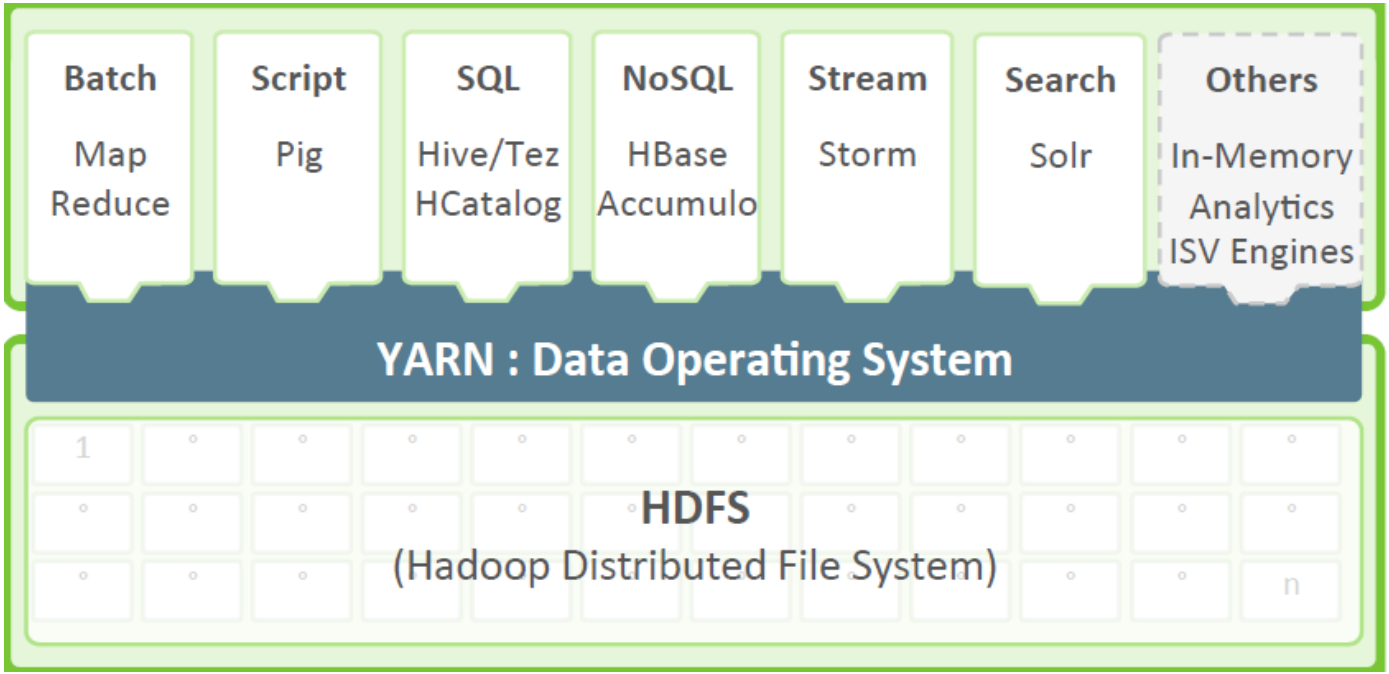
Hadoop evolved directly from commodity scientific supercomputing clusters developed in the 1990s.

Hadoop consists of a parallel execution framework called

- Map/Reduce and
- Hadoop Distributed File System (HDFS).



Latest Developments



HDFS

- Very high fault tolerance
- Can not be updated but corrections can be appended
- File blocks are replicated multiple types

Three types nodes:

Name Node (Directory)

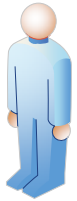
Backup Node (checkpoint)

Data Node-actual data

MapReduce

- A programing framework for library and runtime. just like .NET
- **Map Function -** *Take a task and break it down into small tasks*
- **Reduce Function -** *Combine the partial answers and find the combined list*
- **Master (Job Tracker)**
  - Is where you submit a query. Manages the Task Trackers which do the actual Map or Reduce task.
- **Workers (Task Trackers)**
  - Do the work, just as each nodes in the cluster have a data node, they also have a task tracker

Basics of MapReduce



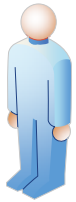
= 400 Seconds

400 bills

1 bill/ sec

Basics of MapReduce

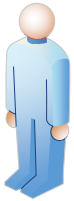
200 Bills



1 bill/ sec

= 200 Seconds

200 Bills

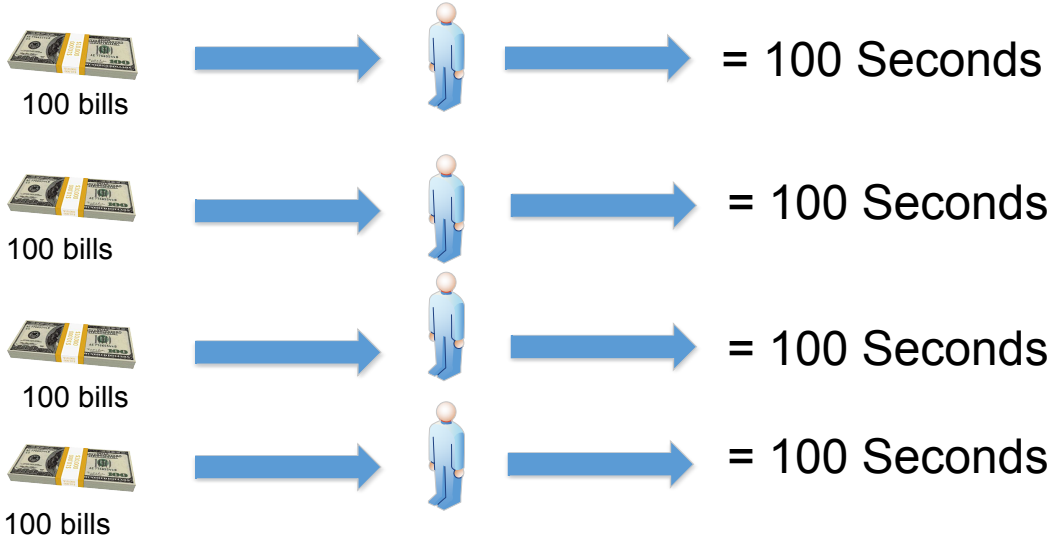


1 bill/ sec

= 200 Seconds

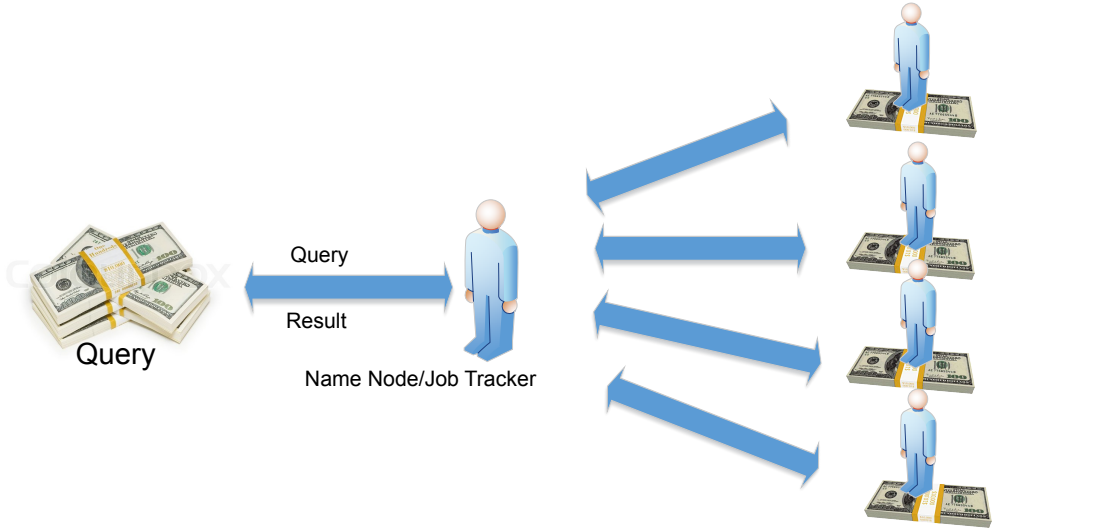
Total =200 seconds

Basics if MapReduce



Total =100 seconds

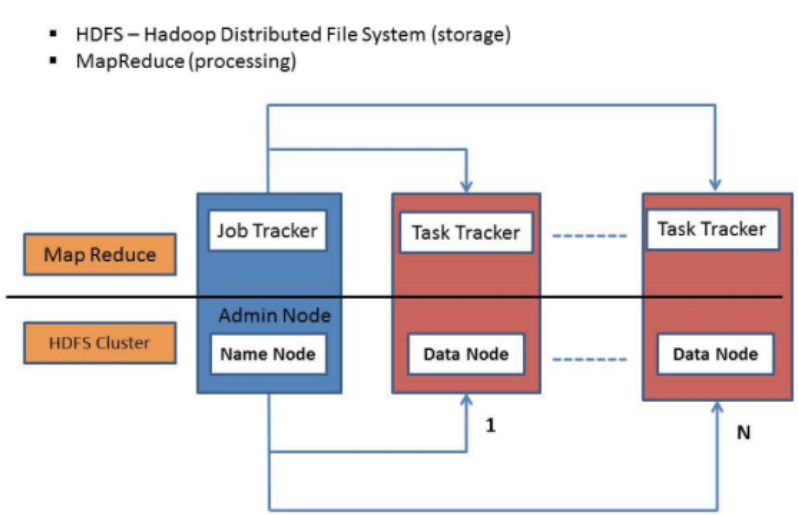
Basics of MapReduce



HDFS and MapReduce

The Main Node: runs the Job tracker and The name node controls the files.

Each node runs two processes: Task Tracker and Data Node



Hive and Pig

MapReduce

- Java
- write many lines of code

Pig

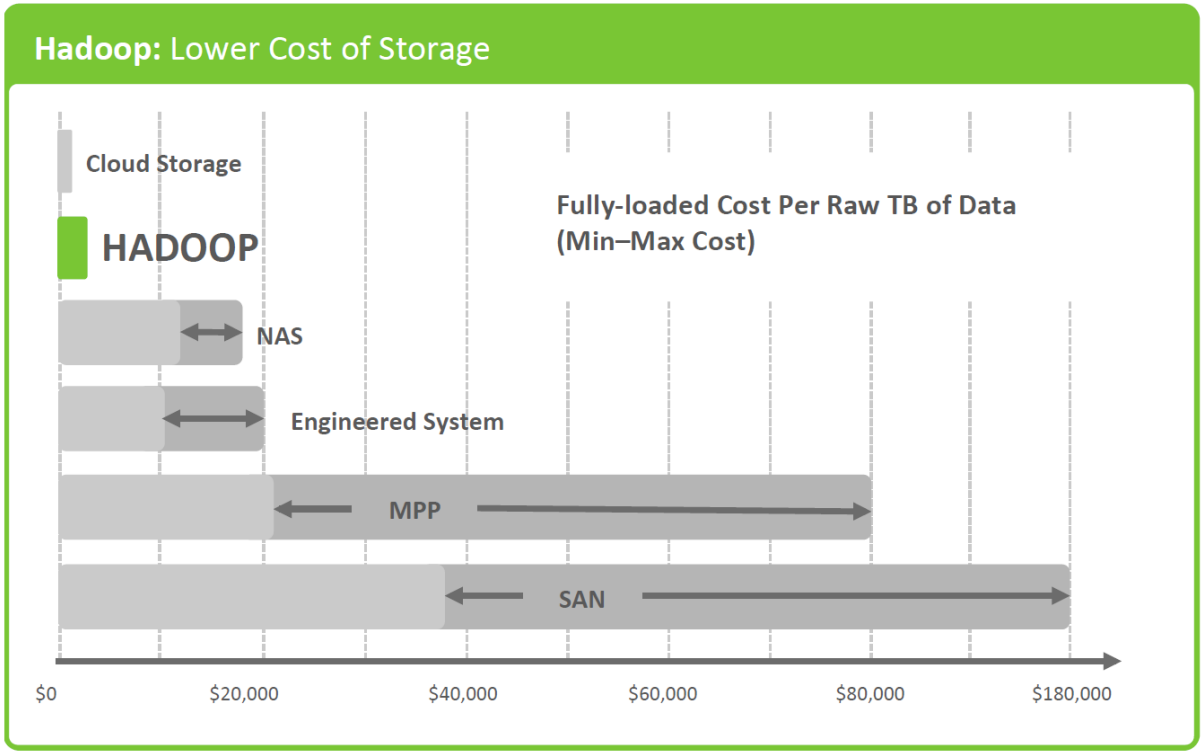
- Mostly used by yahoo
- highly used for data processing
- Shares some constructs with SQL e.g. filtering, selecting, grouping, and ordering. But syntax is very different from sql.
- Is more Verbose
- Needs a lot of training for users with limited procedural programming background.
- Gives you more control over the flow of data.

Hive

- Mostly used by Facebook for analytic purposes
- Used for analytics
- Relatively easier for developers with SQL experience.
- Less control over optimization of data flows compared to Pig

- Not as efficient as MapReduce
- Higher productivity for data scientists and developers

HDFS



* A Modern Data Architecture with Apache Hadoop, Hortonworks Inc. 2014

Big Data Solution Deployment Models

Major Players in Big Data

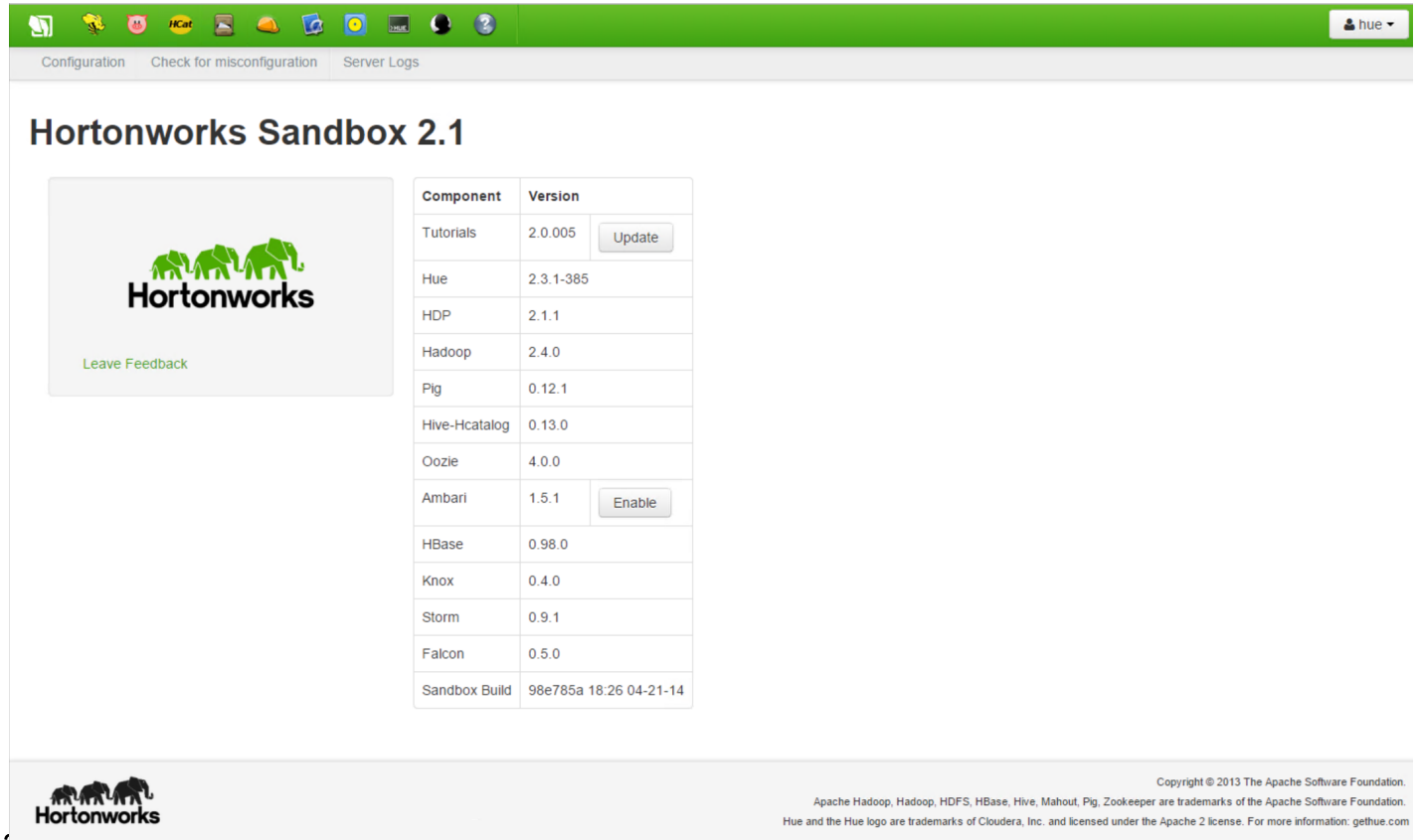
Apache Foundation

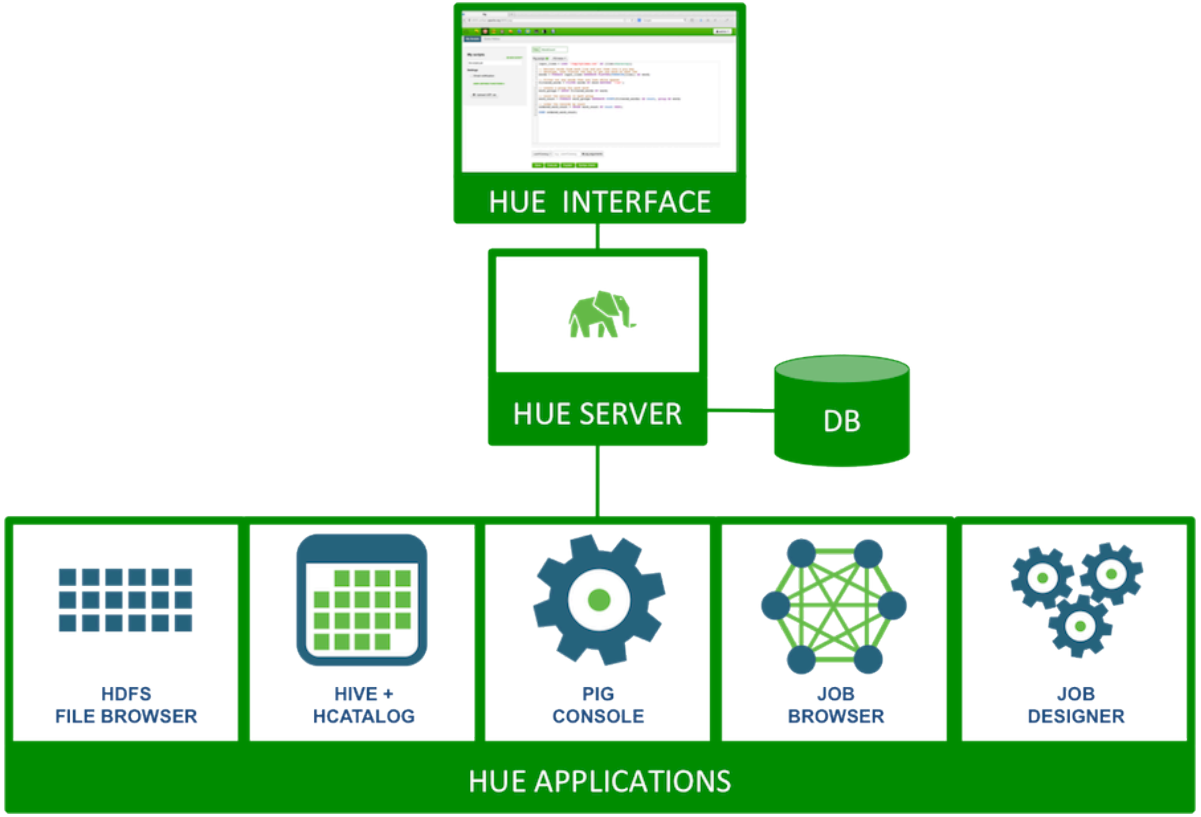
- **Hortonworks**
- **Cloudera**
- **MapR**
- **Pentaho**
- **Amazon (AWS)**
- **…**

Hortonworks

- June 2011 funded by $23 million from Yahoo! and Benchmark Capital as an independent company
- Horton the Elephant - Horton Hears a Who!
- Employs contributors to project Apache Hadoop
- October 2011 partnered with Microsoft : Azure and Windows Server .
- Cloudera founded in October 2008…started the effort to be Microsoft Azure Certified in October 2014.

HDP User Interface



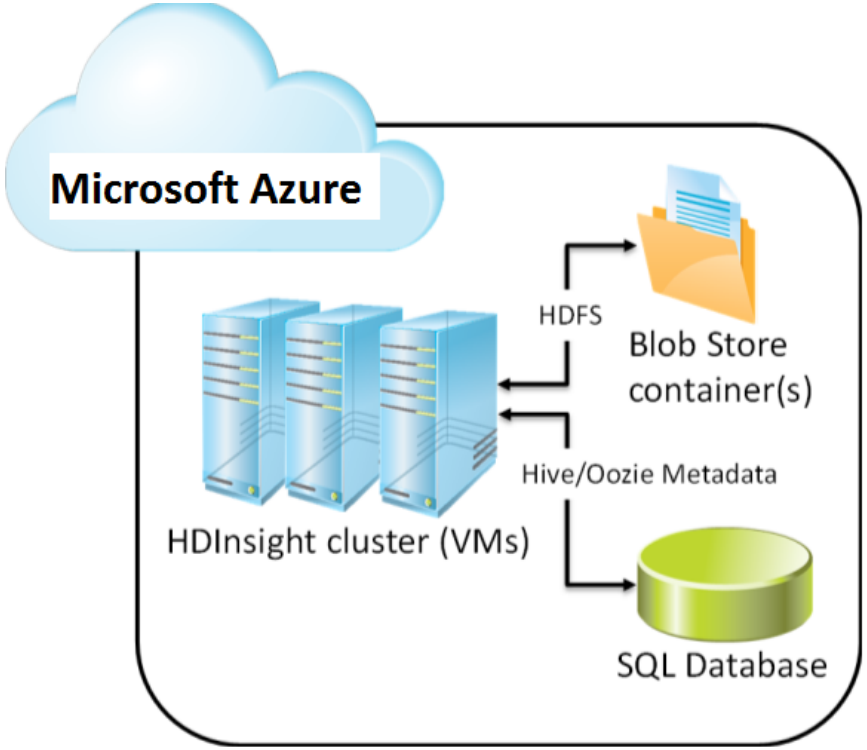


Deployment Models

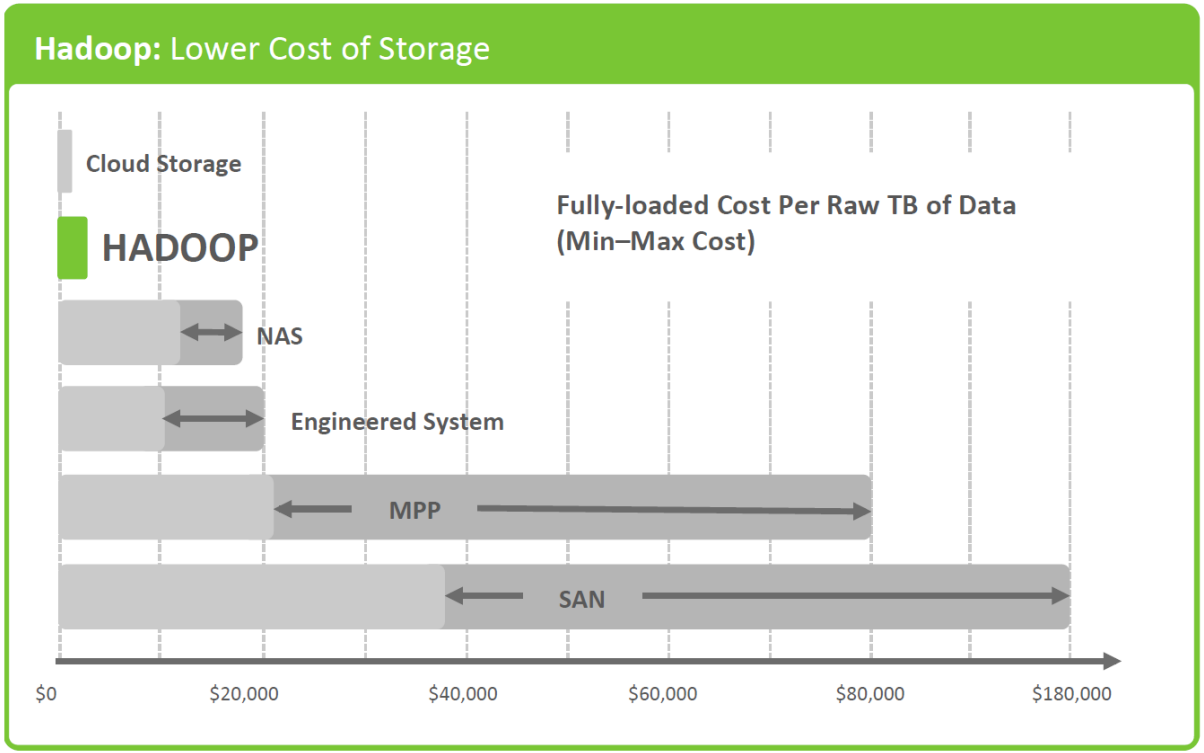
- On Premise Deployment
  - Microsoft Analytics Platform System (APS)
  - Oracle Big Data Appliance
  - Hortonworks Data Platform (HDP)
  - Cloudera's CDH
  - Pivotal Data Computing Appliance (DCA)

- Big Data as a service
  - HDInsight
  - Cloudera on AWS
  - Amazon RedShift
  - Amazon Elastic MapReduce

HDInsight: Hadoop As A Cloud Service

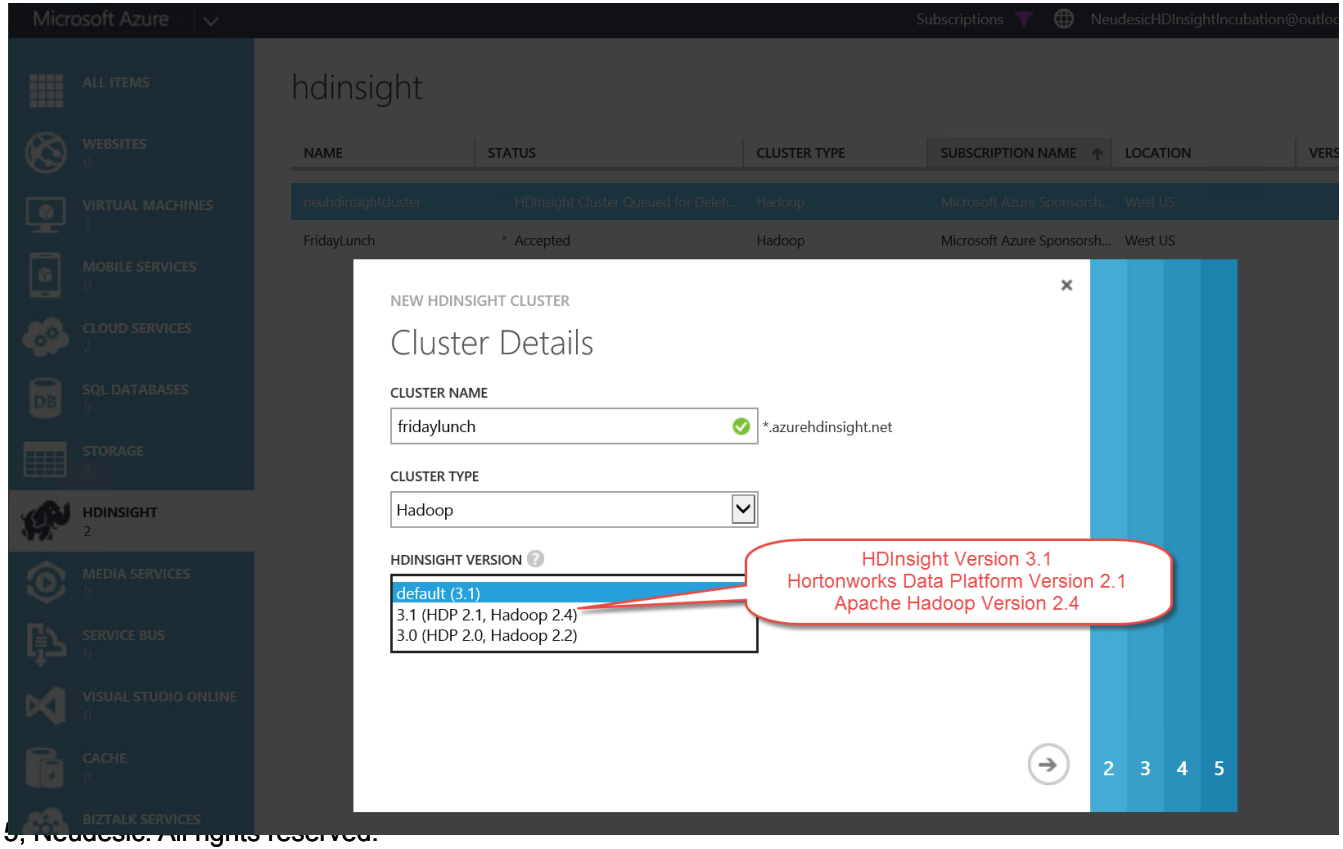


HDFS



* A Modern Data Architecture with Apache Hadoop, Hortonworks Inc. 2014

HDInsight Versions



Architecting the Modern Data Warehouse

The ETL Automation Model

Hadoop: Data Warehouse Workload Optimization

Current Reality

EDW at capacity: some usage from low value workloads

Older transformed data archived, unavailable for ongoing exploration

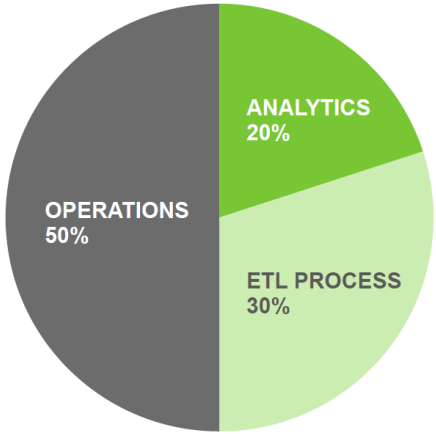
Source data often discarded

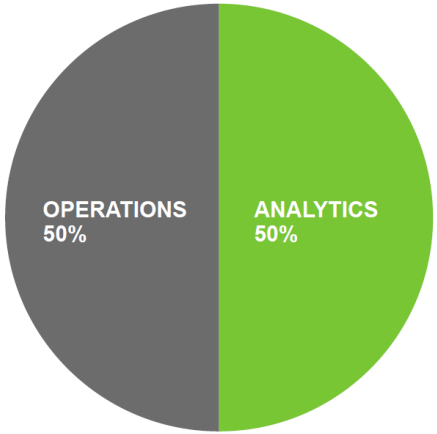
Augment with Hadoop

Free up EDW resources from low value tasks

Keep 100% of source data and historical data for ongoing exploration

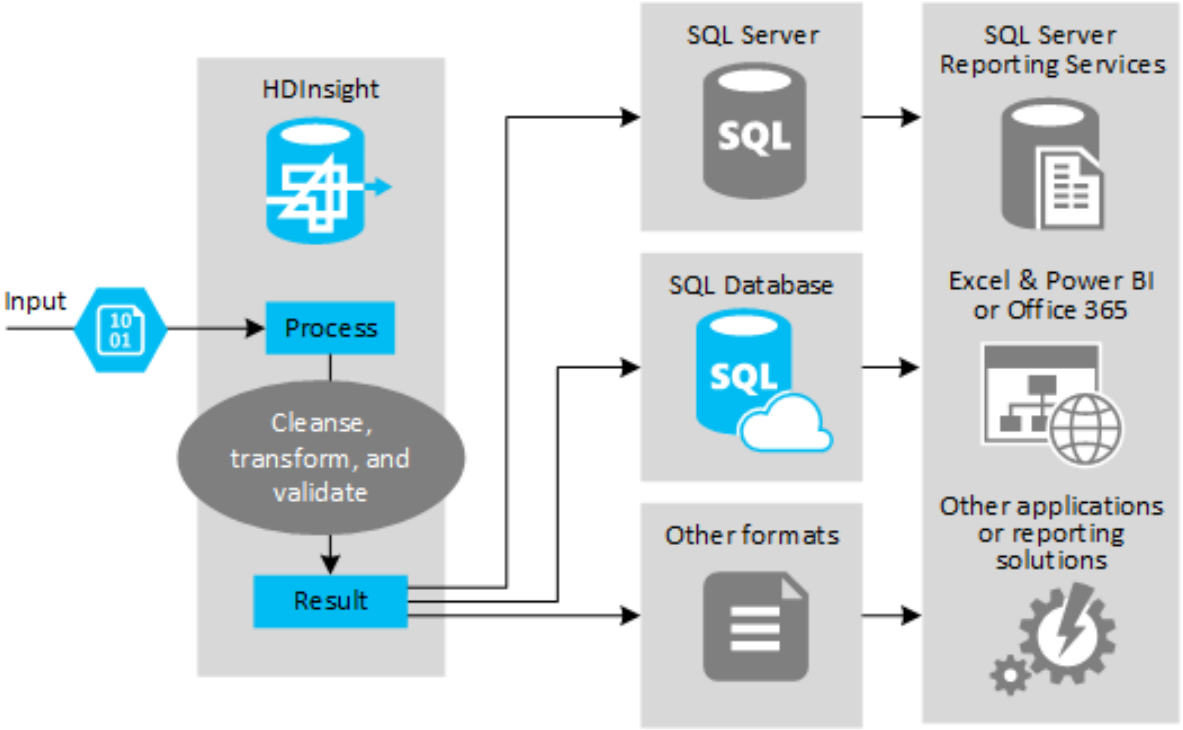
Mine data for value after loading it because of schema-on-read



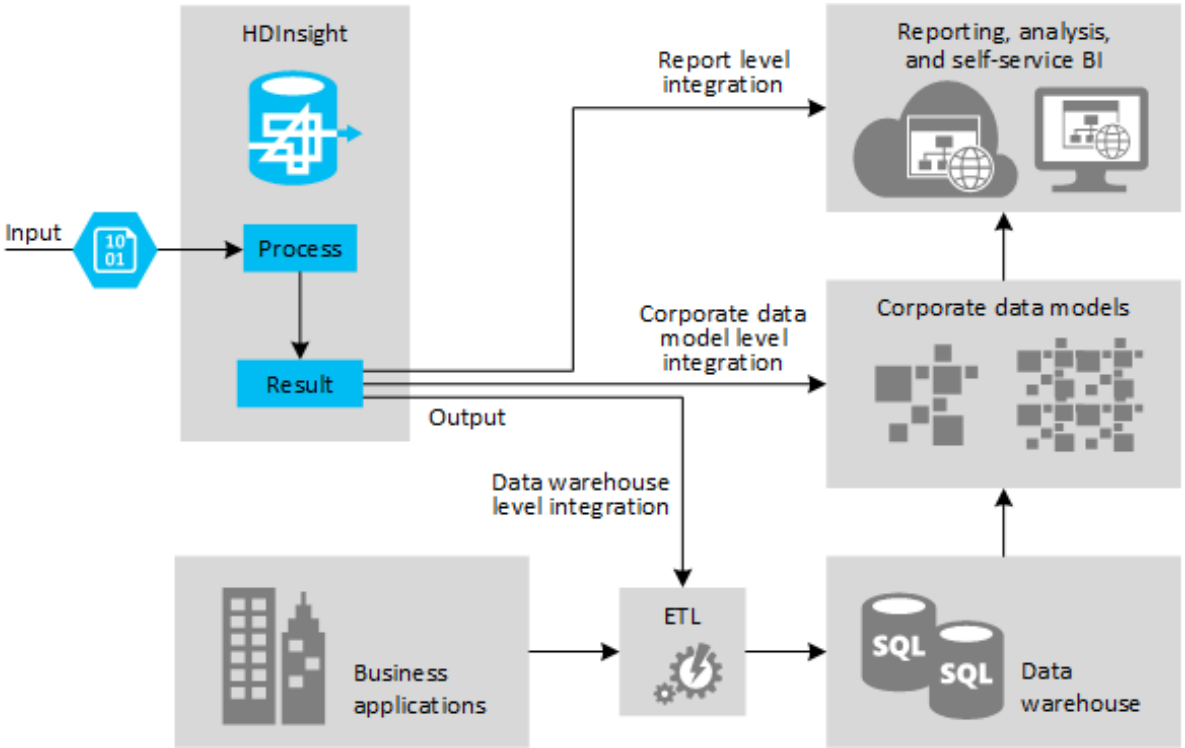


* A Modern Data Architecture with Apache Hadoop, Hortonworks Inc. 2014

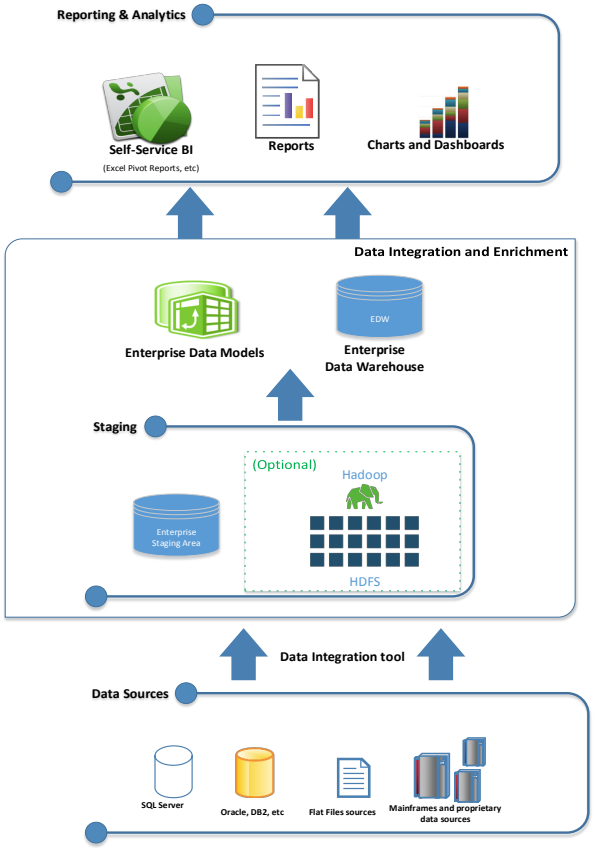
The ETL Automation Model



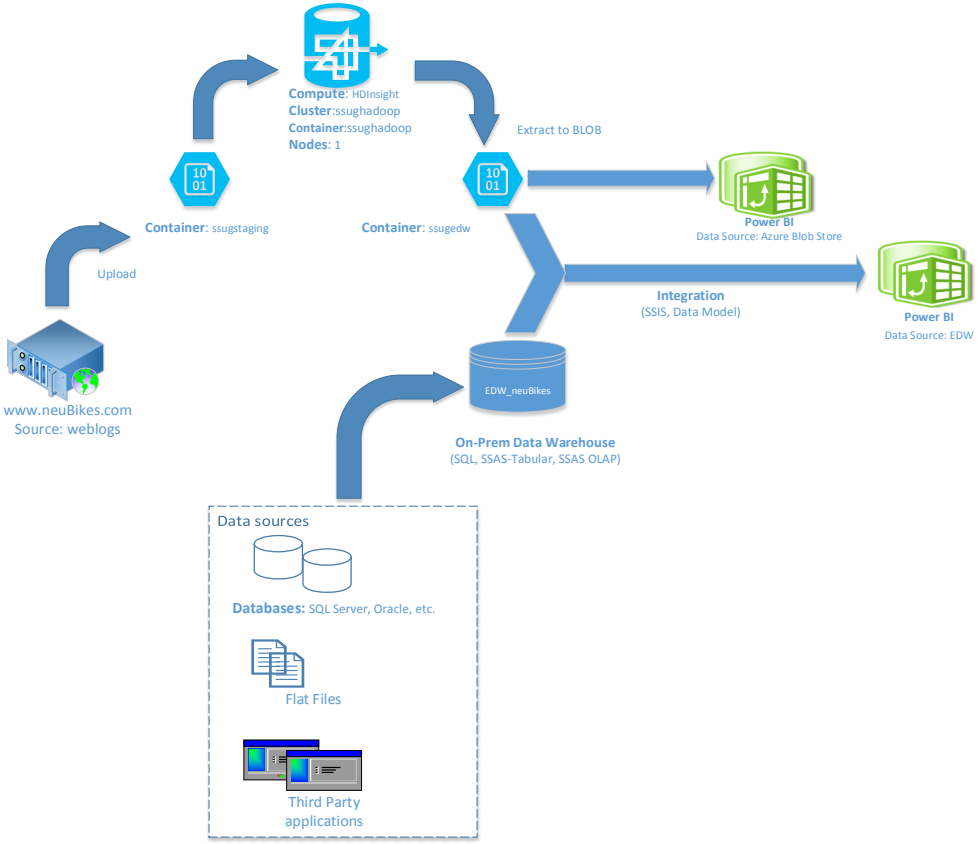
BI-Integration Model



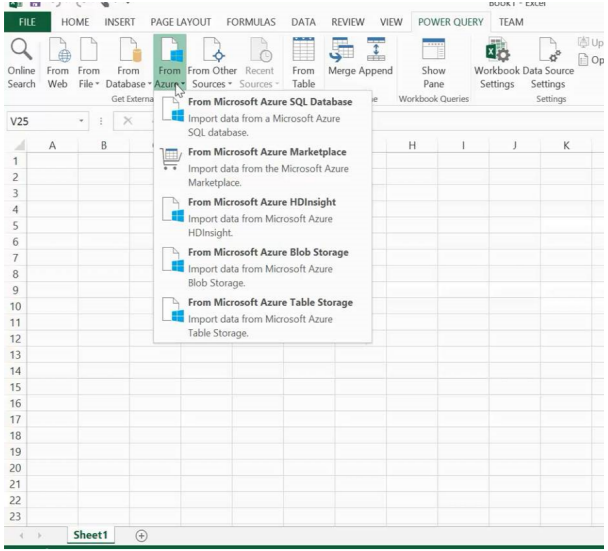
Hybrid BI-Integration Model

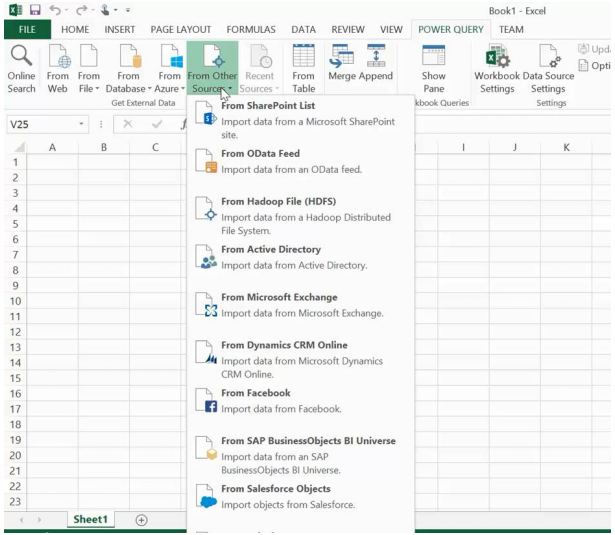


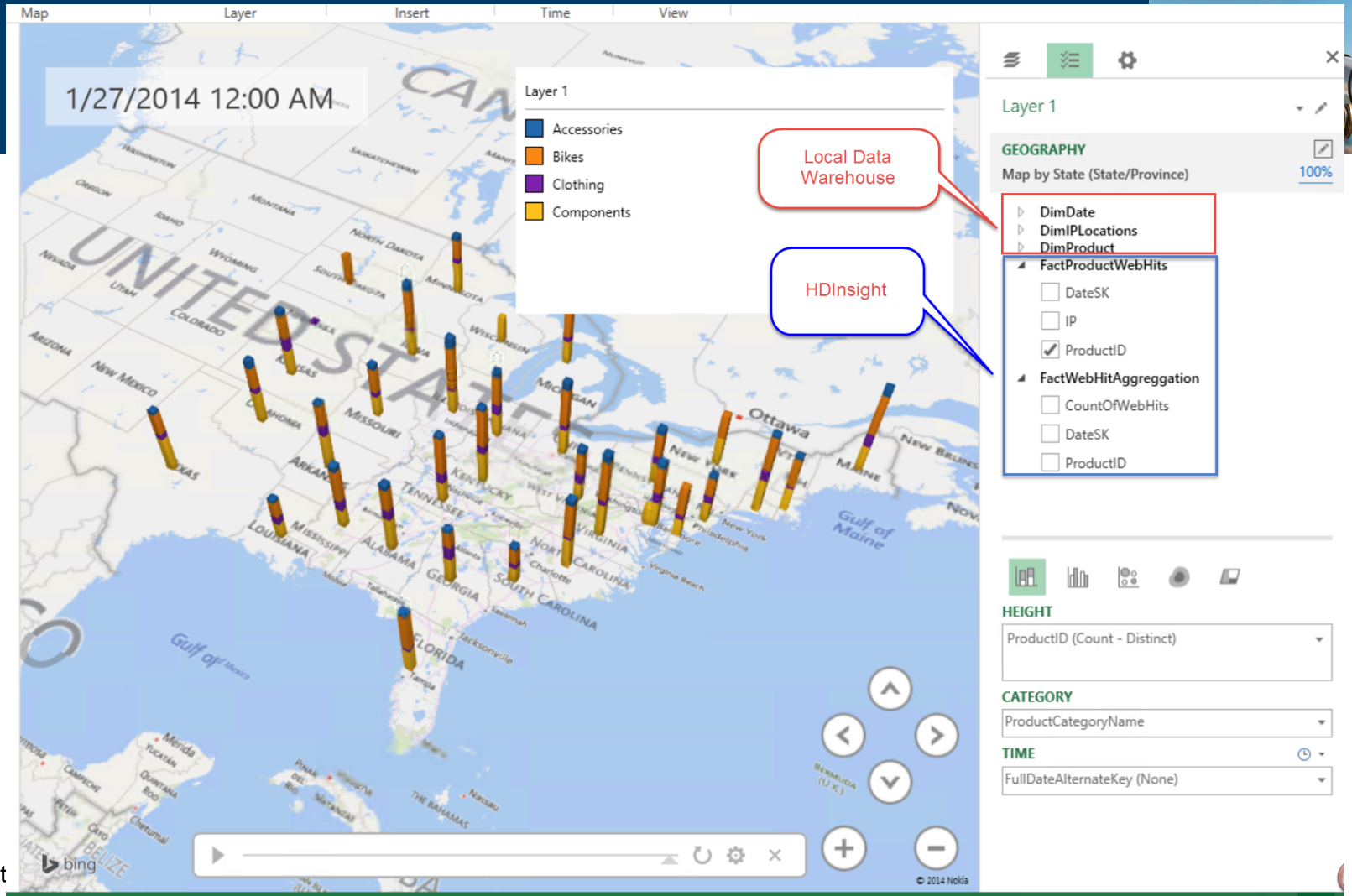
Hybrid BI-Integration Model



Hybrid BI-Integration Model






Map
Layer
Insert
Time
View
1/27/2014 12:00 AM
Layer 1
Accessories
Bikes
Clothing
Components
Local Data Warehouse
HDInsight
Layer 1
GEOGRAPHY
Map by State (State/Province)
100%
DimDate
DimIPLocations
DimProduct
FactProductWebHits
DateSK
IP
ProductID
FactWebHitAggreggation
CountOfWebHits
DateSK
ProductID
HEIGHT
ProductID (Count - Distinct)
CATEGORY
ProductCategoryName
TIME
FullDateAlternateKey (None)
bing
© 2014 Nokia

Summary

Summary

- Understand your data growth to determine when to “Scale-Out”.
- Determine the right tool for the workload you have.
- Choose the right deployment of Big Data Solutions
- Hybridize, do not start from scratch!

Questions?

Questions and Discussion