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





The Inside Scoop on Hadoop

Topics Covered



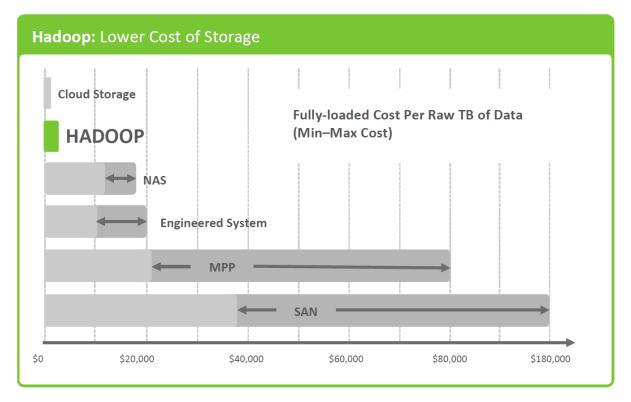
- Understanding Hadoop
- Big Data Solution Deployment Models
- Architecting the Modern Data Warehouse
- Summary



Understanding Hadoop

Big Data = Hadoop?





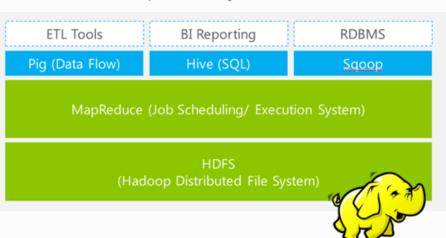
The Fundamentals of Hadoop



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

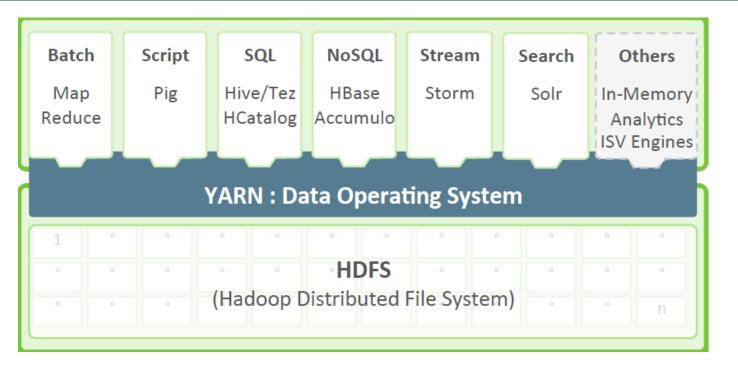
Hadoop consists of a parallel execution framework called

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



The Hadoop Ecosystem

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)

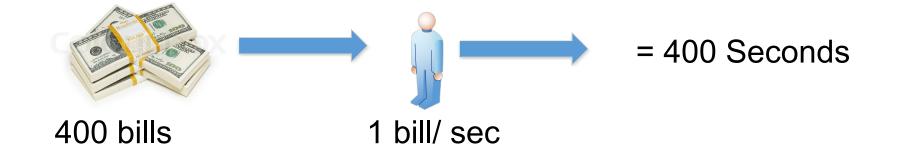
 Is where you submit a query. Manages the Task Trackers which do the actual Map or Reduce task.

Workers (Task Trackers)

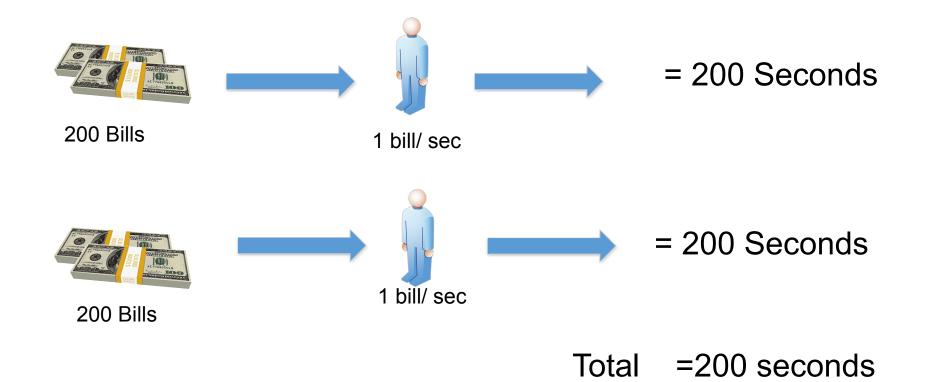
 Do the work, just as each nodes in the cluster have a data node, they also have a task tracker

Basics of MapReduce



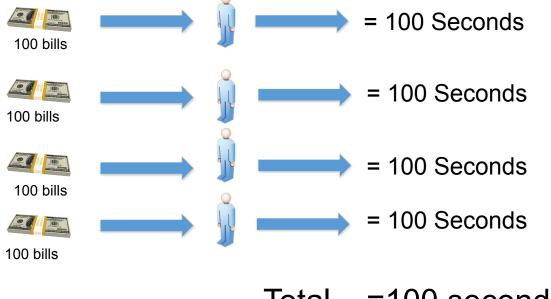


Basics of MapReduce



Basics if MapReduce

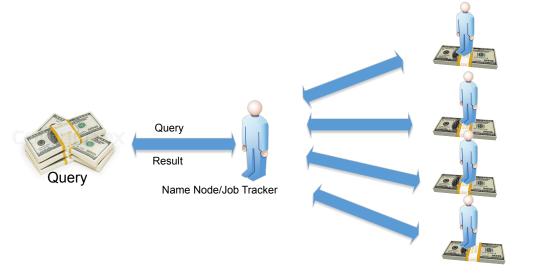




Total =100 seconds

Basics of MapReduce





Data Nodes/Task Trackers

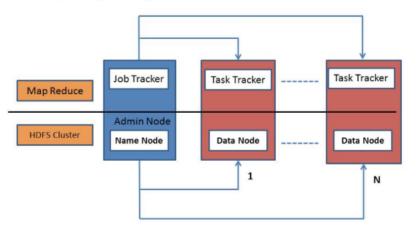
HDFS and MapReduce



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

Each node runs two processes: Task Tracker and Data Node

- HDFS Hadoop Distributed File System (storage)
- MapReduce (processing)



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.

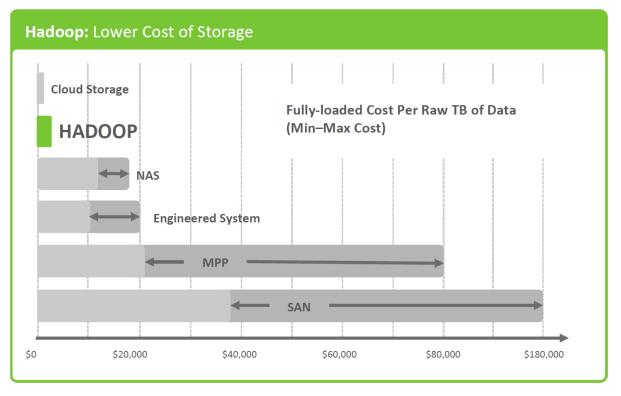
Hive

- 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





© Copyright 2015, Neudesic. All rights reserved.

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



Big Data Solution Deployment Models

Major Players in Big Data





- 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

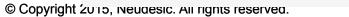
5	Â,	8	HCat		۲		$\overline{\mathbf{O}}$	HUE	۹	3	
Confi	iguratio	n C	Check fo	or misc	onfigura	ation	Serve	er Logs			

Hortonworks Sandbox 2.1

	Component	Version
	Tutorials	2.0.005 Update
Hortonworks	Hue	2.3.1-385
Hortonworks	HDP	2.1.1
Leave Feedback	Hadoop	2.4.0
	Pig	0.12.1
	Hive-Hcatalog	0.13.0
	Oozie	4.0.0
	Ambari	1.5.1 Enable
	HBase	0.98.0
	Knox	0.4.0
	Storm	0.9.1
	Falcon	0.5.0
	Sandbox Build	98e785a 18:26 04-21-1



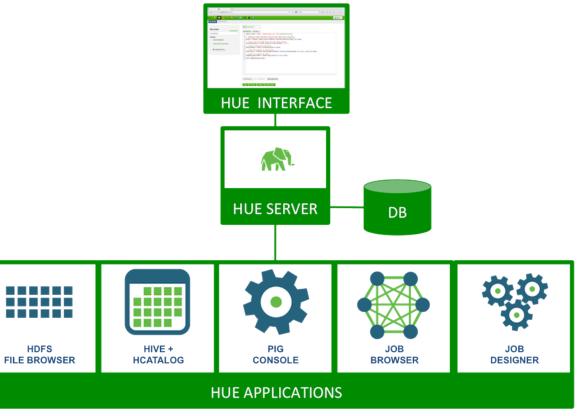
Copyright 9 2013 The Apache Software Foundation. Apache Hadoop, Hadoop, HDFS, HBase, Hive, Mahout, Pig, Zookeeper are trademarks of the Apache Software Foundation. Hue and the Hue logo are trademarks of Cloudera, Inc. and licensed under the Apache 2 locense. For more information: gethue.com





💄 hue 👻





Deployment Models

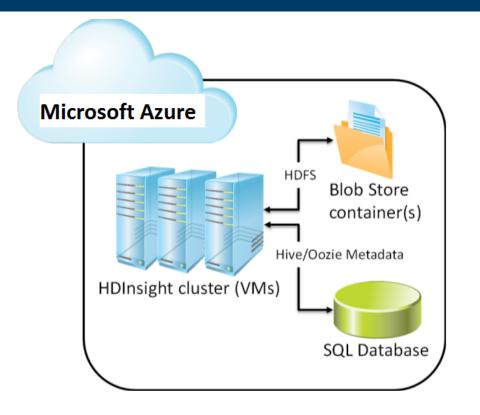
On Premise Deployment

- 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



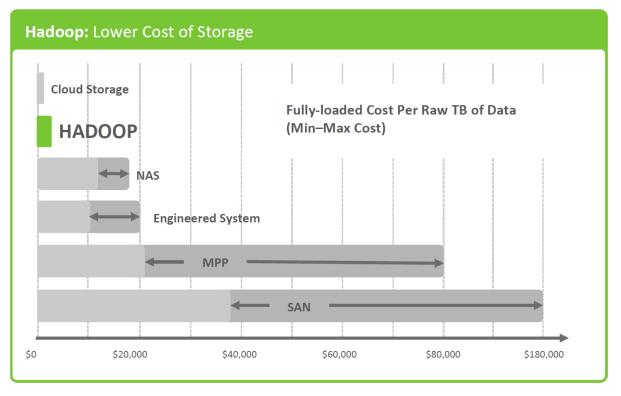
- Microsoft

Azure

Starting up...

HDFS





© Copyright 2015, Neudesic. All rights reserved.

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

HDInsight Versions

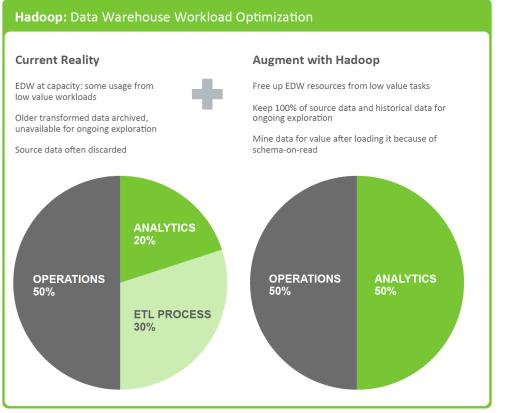


Micro	osoft Azure 🛛 🗸				Subscriptions 🍸 🌐 N	leudesicHDInsi	ghtIncubation@ou	tloc
		hdinsight						
		NAME	STATUS	CLUSTER TYPE	SUBSCRIPTION NAME	LOCATION	v	/ER
6		FridayLunch	Accepted SIGHT CLUSTER	Hadoop	Microsoft Azure Sponsorsh	n West US		
-69-			er Details					
DB		CLUSTER NA		*.azurehdinsight.net				
		CLUSTER TY		•.azurehdinsight.net				
(A)	HDINSIGHT	Hadoop		~				
٢		HDINSIGHT	1)	Hortonworks Da	ght Version 3.1 ata Platform Version adoop Version 2.4	2.1		
ĘΣ			2.1, Hadoop 2.4) 2.0, Hadoop 2.2)					
M								
F					(\rightarrow)	2 3 4	5	



Architecting the Modern Data Warehouse

The ETL Automation Model

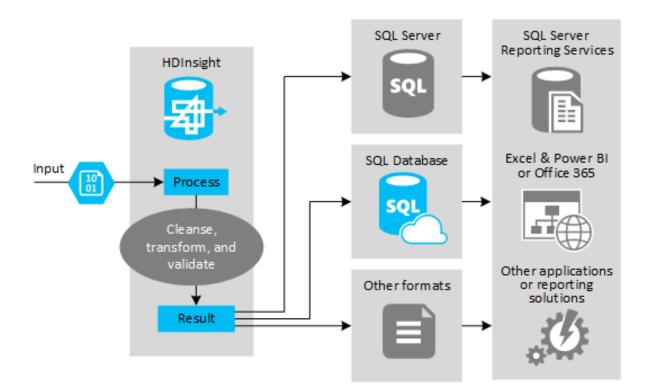


© Copyright 2015, Neudesic. All rights reserved.

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

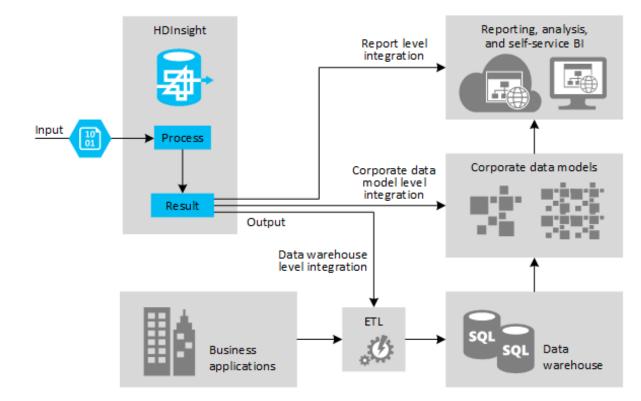
The ETL Automation Model





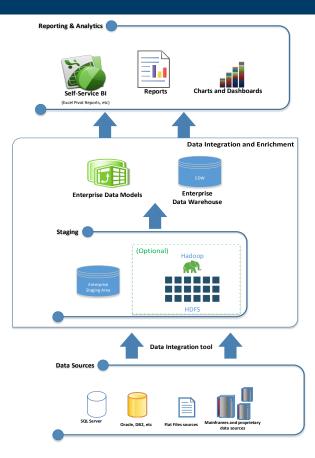
BI-Integration Model





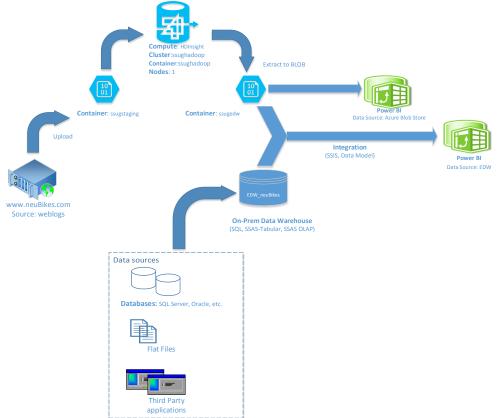
Hybrid BI-Integration Model





Hybrid BI-Integration Model

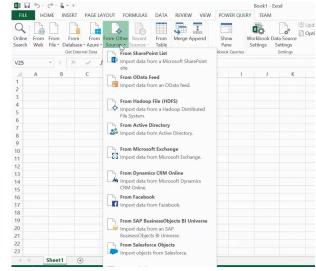




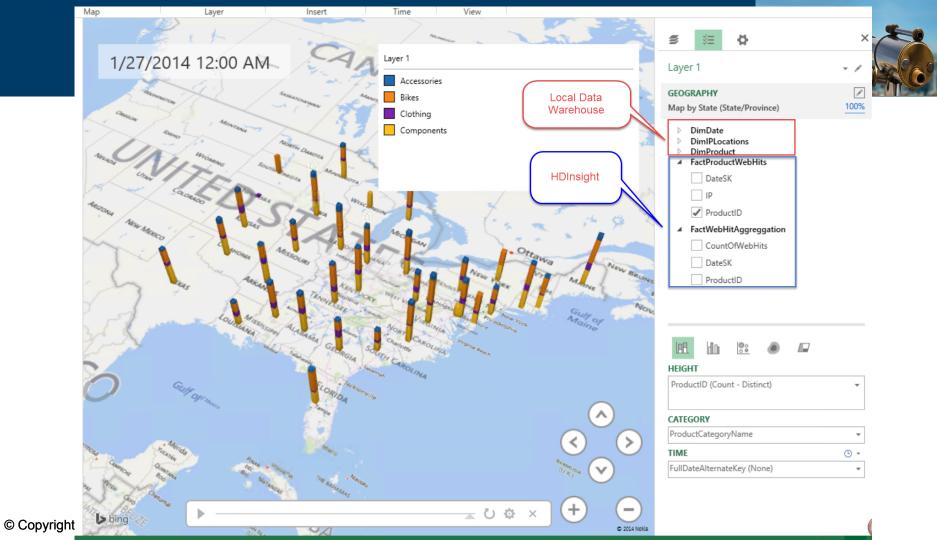
Hybrid BI-Integration Model



FILE	The second	IME I	NSERT	PAGE L	1001	FORMULAS	DATA	REVIEW	VIEV	100	VER QUEF	TEAM		
Online iearch	From	From File *	From Database			her Recent	From Table	Merge A	Append	Sho		Workbook Settings	Data Source Settings	🕼 Up 📄 Op
			Get Extern			icrosoft Azu			e	Workbook	Queries		Settings	
V25		* 1	×		Import of SQL data	lata from a M abase.	icrosoft A	zure						
1	А	B		1	From Mi	icrosoft Azu	e Marke	place		Н	1	J	K	
1						lata from the								
2					Marketp		Microsof	Pacure						
3					From Mi	icrosoft Azu		abt						
4						lata from Mic								
5					HDInsig		0301174	are .						
6						icrosoft Azur	Plah C							
7				- 17		lata from Mic								
8				_	Blob Sto		IOSOIT AZ	ne						
9						icrosoft Azu								
10				1		lata from Mic								
11				-	Table Sto		rosoit Az	ne						
12					Tuble 50	Juge,			-					
13														
14														
15														
16														
17														
18														
19														
20														
21														
22														
23			(+)											



*		
Connect	Search	_
Excel Fact fire Access Statistical file Cher Files Dia a server Tableau Server Microsoft SQL Server MySGL Oracle Amazon Redshift More Servers CountERPart (SAP, Fil) Sales Overview (SNYOC. Sample - SuperSone World Indicators	Actian Vectorwise Amazon Rechift Anter Database Amazon KM Cioucera Hadoop DataBas Enterprise EXSolution Firebrid Google Analytics Google BigCaery Hettowork Hadoop Hive H [®] Vertica IBM BigLaghts HBM BigLaghts IBM Merz23 Maph Hadoop Hive MarkLogic	





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 and Discussion