International Journal of Engineering Science Invention (IJESI)
ISSN (Online): 2319 – 6734, ISSN (Print): 2319 – 6726
www.ijesi.org ||Volume 7 Issue 3 Ver. 4 || March 2018 || PP05-08

Big Data

Sayalishete
Lecturer, Bgit

Abstract: *As We Transact With Tera And Peta Bytes Of Data In Our Daily Life, We Never Analyze The Impact That Data Holds For The User And For The Clients And Companies Who Deal With These Data. The Growing Data Is An Issue Of Concern For The Application Development Firms To Store The Data. The Data Is Required As Many Applications Run On These Data With Security Attached To These Data. The Large Growing Data Termed As Big Data To The Information Technology And The Software Development Industry. The Large Data Needs To Be Handled, And Becomes Easy To Handle If It Is Divided Into Clusters. Hadoop Is A Framework For Running Applications On Large Clusters. In This Paper We Discuss And Analyze The Output On The Basis Of Case Study To Compare Various Infrastructure To Handle Big Data With Horton Works Sandbox, Pig, Hive, Hcatalogand Jaql And Also Refereeing To The Concept Of Cloudera In Brief.*

Keywords*: Big Data, Hadoop, Clusters, Pig, Hive, Ibm Infosight.*

Date of Submission: 27-02-2018 Date of acceptance 15-03-2018

I. Introduction

The Analysis Of Big Data By Dividing The Tasks Into Various Clusters Become Easy To Work On. Hadoop Implements The Principle Of Computational Strategy Known As Map/ Reduce. The Application Works On The Various Nodes In The Cluster. Hadoop Provides A Distributed File System Known As Hdfs I.E. Hadoop Distributed File System That Stores The Data On The Computational Node, Providing Very High Aggregate Bandwidth Across The Cluster. The File Structure And The Principle Computational Using Map/ Reduce Are Designed So That Node Failures Are Automatically Handled By The Framework Structure. The Basic Structure Of Hadoop Architecture Is Shown In Figure 1.

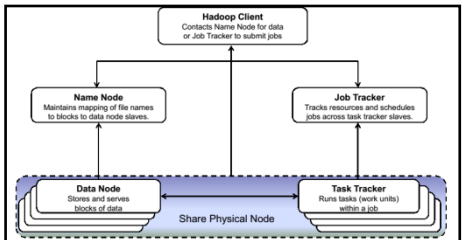

Figure 1: The Basic Hadoop Architecture.

The Hdfs Is The Foundation For Forming The Hadoop Clusters With Various Components As Shown In Figure 2.

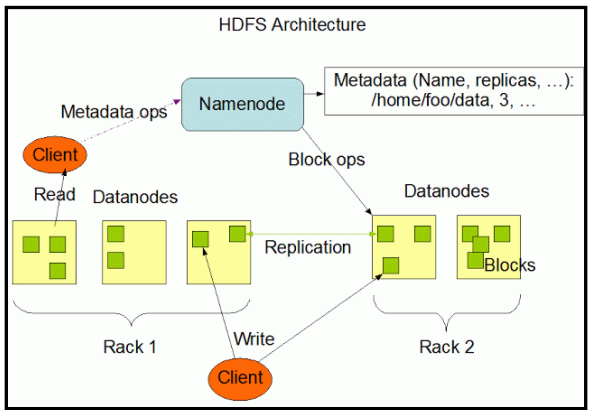

Figure 2: The Hdfs Architecture

.

The File Structure Manages How Data Is Stored In The Hadoop Cluster And Is Also Responsible For Distributing The Data Across The Data Nodes, Managing Replication For Redundancy And Administrative Tasks Like Adding, Removing And Recovery Of Data Nodes.

Hadoop In Hortonworks Sandbox

1. Once The Virtual Box Is Completely Installed.
2. Go To Local Host 127.168.0.0:8888, To Go To Hortonworks Sandbox And Click On Go To Sandbox As Shown In Figure 3.

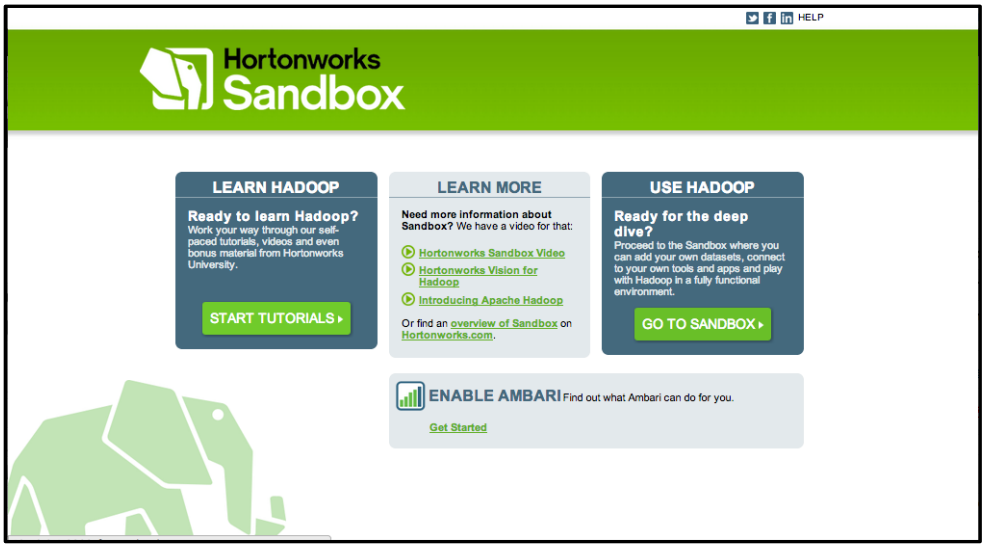

Figure3:HortonworkSandbo

Data Processing With Pig

Pig Is A High Level Scripting Language That Is Used With Apache Hadoop. Pig Excels At Describing Data Analysis Problems As Data Flows. Pig Is Complete In That You Can Do All The Required Data Manipulations In Apache Hadoop With Pig. Pig Can Ingest Data From Files, Streams Or Other Sources Using The User Defined Functions (Udf).

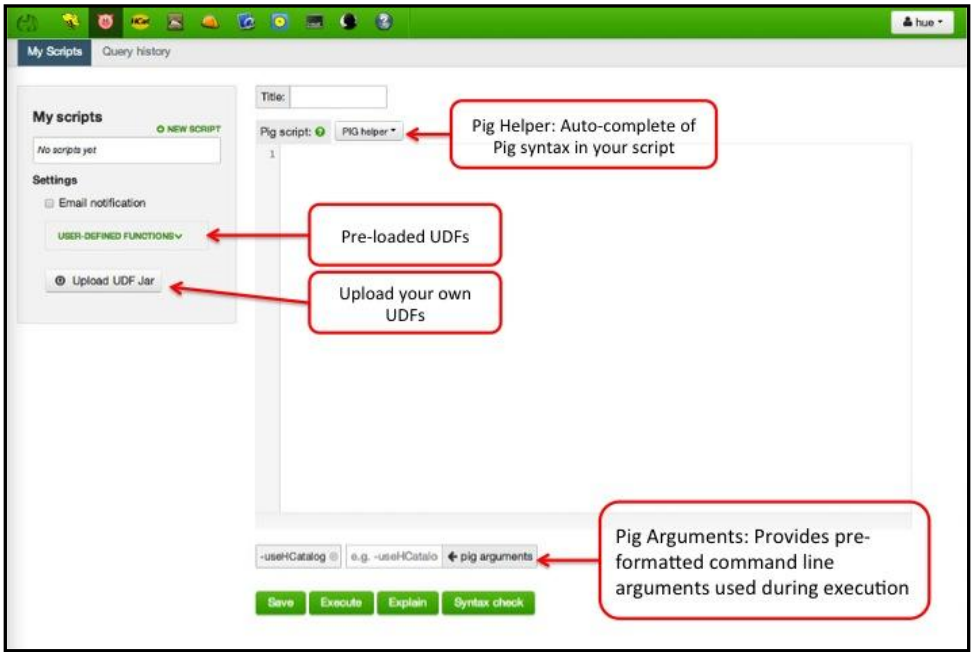

Figure4: ScreeOfPig

Pig Is A Language For Expressing Data Analysis And Infrastructure Processes. Pig Is Translated Into A Series Of Mapreduce Jobs That Are Run By The Hadoop Cluster. Pig Is Extensible Through User-Defined Functions That Can Be Written In Java And Other Languages. Pigscripts Provide A Highlevel Language To Create The Mapreduce Jobs Needed To Process Data In A Hadoop Cluster.

Data Processing With Hive

Hive Is A Component Of Hortonworks Data Platform (Hdp). Hive Provides A Sql-Like Interface To Data Stored In Hdp. In The Previous Tutorial We Used Pig Which Is A Scripting Language With A Focus On Data Flows. Hive Provides A Database Query Interface To Apache Hadoop.

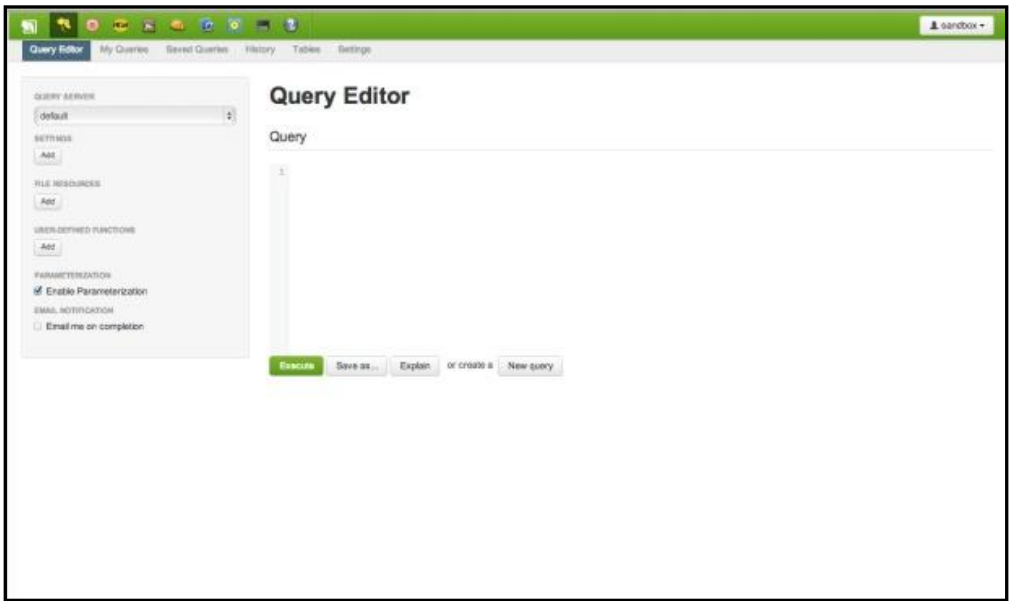

Figure5:QueryEditorHive

The Apache Hive Project Provides A Data Warehouse View Of The Data In Hdfs. Using A Sql-Like Language Hive Lets You Create Summarizations Of Your Data, Perform Ad-Hoc Queries, And Analysis Of Large Datasets In The Hadoop Cluster. The Overall Approach With Hive Is To Project A Table Structure On The Dataset And Then Manipulate It With Hiveql. Since You Are Using Data In Hdfs Your Operations Can Be Scaled Across All The Data Nodes And You Can Manipulate Huge Datasets.

Data Processing With Hcatalog

The Function Of Hcatalog Is To Hold Location And Metadata About The Data In A Hadoop Cluster. This Allows Scripts And Mapreduce Jobs To Be Decoupled From Data Location And Metadata Like The Schema. Additionally Since Hcatalog Supports Many Tools, Like Hive And Pig, The Location And Metadata Can Be Shared Between Tools. Using The Open Apis Of Hcatalog Other Tools Like Teradata Aster Can Also Use The Location And Metadata In Hcatalog. In The Tutorials We Will See How We Can Now Reference Data By Name And We Can Inherit The Location And Metadata.

Figure6: TablesOfHcat

Data Processing With Jaql

The Query Language Adopted By The Ibm Big Data Ecosystem For Java Script Objet Notation (Json). It Supports Both Structured And Non-Structured Queries And Allows You To Select, Join, Group, And Filter Data That Is Stored In Hdfs.

1. Jaql's Query Language Was Inspired By Other Programming And Query Languages, Including Lisp, Sql, Xquery, And Pig
2. Jaql Is A Functional, Declarative Query Language That Is Designed To Process Large Data Sets.
3. For Parallelism, Jaql Rewrites High-Level Queries, As Appropriate, Into "Low-Level" Queries Consisting Of Mapreduce Jobs.

The Importance Of Jaql Is

1. The Real Beauty Of The Language Is That It Transparently Exploits Massive Parallelism Using Apache Hadoop's Mapreduce Processing.
2. Jaql Code Can Be Packaged In A Way That Fosters Reuse. It Is Extensible, Allowing Functions Written In Other Languages, For Example Java, To Be Invoked From Within Your Jaql Code.

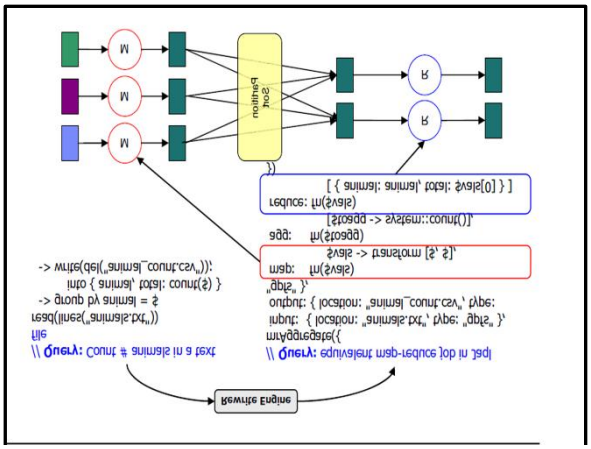

Figure 7: Jaql Structure

II. Conclusion

Hadoop And The Data Warehouse Will Often Work Together In A Single Information Supply Chain. When It Comes To Big Data, Hadoop Excels In Handling Raw, Unstructured And Complex Data With Vast Programming Flexibility.

Data Warehouses Also Manage Big Structured Data, Integrating Subject Areas And Providing Interactive Performance Through Bi Tools. It Is Rapidly Becoming A Symbiotic Relationship. Some Differences Are Clear And Identifying Workloads Or Data That Runs Best On One Or The Other Will Be Dependent On Your Organization And Use Cases. As With All Platform Selections, Careful Analysis Of The Business And Technical Requirements Should Be Done Before Platform Selection To Ensure The Best Outcome. Having Both Hadoop And A Data Warehouse Onsite Greatly Helps Everyone Learn When To Use The Priority Of Hadoop With Data Warehouse.

Reference

[1]. Dean, J. And Ghemawat, S., "Mapreduce: Simplified Data Processing On Large Clusters."Appeared In Proceedings Of The Sixth Symposium On Operating System Design And Implementation, San Francisco, Ca, December, 2004.Available Online At Http://Labs.Google.Com/Papers/Mapreduce.Html, March 2010.

[2]. Apache Nutch Project, Http://Lucene. Apache .Org/Nutch/, March 2010.

[3]. Apache Hadoop Project, Http://Hadoop. Apache .Org/, March 2010.

[4]. Patterson, J., "The Smart Grid: Hadoop At The Tennessee Valley Authority (Tva)," Cloudera Blog, Http://Www.Cloudera.Com/ Blog/2009/06/Smart-Grid-Hadoop-Tennessee-Valley -Authority-Tva/, June 2, 2009.

[5]. White, T., Hadoop: The Definitive Guide. O'reilly Media, May 2009.

[6]. Lin, J. And Dyer, C., Data-Intensive Text Processing With Mapreduce. To Be Published By Morgan And Claypool. Pre-Press Edition Available At Http://Www.Umiacs.Umd.Edu/~ Jimmylin/Mapreduce-Book-20100219.Pdf, February 2010.

[7]. Segaran, T. And Hammerbacher, J., Beautiful Data. O'reilly Media, July 2009.

[8]. Hadoop And The Data Warehouse:When To Use Which,Dr. Amrawadallah,Founder And Cto, Cloudera, Inc.Dangraham,General Manager, Enterprise Systems, Teradata Corporation.

[9]. Http://Www.Tei-C.Org/Index.Xml

[10]. Http://History.State.Gov/Historicaldocuments

[11]. Leetaru, K. (Forthcoming). "Fulltext Geocoding Versus Spatial Metadata For Large Text Archives: Towards A Geographically Enriched Wikipedia", D-Lib Magazine

Sayalishete" Big Data"International Journal of Engineering Science Invention (IJESI), vol. 07, no. 03, 2018, pp05-08