**Big Data and Hadoop**

**POC Project**

**By**

**Abhishek Kumar Jaiswal**

**Email:- <ia.malhotra@gmail.com>**

**Twitter:- @abhishek\_jaisw**

Table Of Contents

|  |  |  |
| --- | --- | --- |
| Sr | Contents | Page |
| 1. | 1. Introduction |  |
| 2. | 2. Problem Statement |  |
| 3. | 3. Solutions |  |
| 4. | 3.1 Map Reduce Solution |  |
| 5. | 3.2 Pig Solution |  |
| 6. | 3.3 Hive Solution |  |
| 7. | 4. Conslusion |  |
| 8. | 5. Reference |  |
|  |  |  |

**1. Introduction**

These Documentation Recorded for Proof Of Concept Project Submission After Attending 3 weeks rigrious training in Edureka to Learn the Basic and Advanced Concept of Big Data and Hadoop

**2. Problem Statement**

To complete Proof of Concept Project we are Provided with Books Crossing Data which Consists of 3 logs files

* BX-Books.csv
* BX-Users.csv
* BX-Book-Ratings.csv

1. BX-Books.csv

This file contents The details about the book Such as

* 1. ISBN
  2. Book Title
  3. Book Author
  4. Year of Publication
  5. 3 Different forms of URL's
     1. image-url-s (small)
     2. image-url-m (medium)
     3. image-url -l (large)

all these url points to amazon website

1. BX-Users.csv

This file contents logs about users

* 1. User-ID
  2. Location of users
  3. Age

1. BX-BooksRatings.csv

This File Contents logs and act as mediatior between Bx-Books and

Bx-Users which contents

* 1. Users-ID (Related to Bx-Users.scv)
  2. ISBN (Related to BX-Books.csv)
  3. Book-rating

So from the Above logs files we have to Find Solution for Following

1. Find out the Frequency of books published each year
2. Find out in which year Maximum Number of books pubished
3. Find out how many book wew published based on ranking in year 2002

3. Solutions

3.1 Map-Reduce

To analyze given book dataset for problem

3.2 Processing With Pig

Apache Pig provides an engine for executing data flows in parallel on Hadoop. It includes a language–Pig Latin–for articulating these parallel data flows. Pig Latin includes operators for many of the traditional data operations such as join, sort, filter, etc., as well as the ability for users to develop their own functions for reading, processing and writing data. Needless to say, Pig runs on Hadoop and it makes use of both the Hadoop Distributed File System [HDFS] and Hadoop’s processing system [MapReduce]. Grunt is Pig’s interactive shell that enables users to enter Pig Latin interactively and provides a shell for users to interact with HDFS.

3.3 Processing With Hive

Hive is a tool in the Hadoop echo system that delivers an SQL language for querying data stored in the Hadoop Distributed file system. Hive is considered friendlier and more familiar to users who are used to using SQL for querying data.

We will use HIVE internal tables and HiveQL to store and perform analytics on data. Tables are called managed or internal tables when Hive controls the lifecycle of their data. Hive stores data for internal tables in a subdirectory under the directory defined by hive.metastore.warehouse.dir. When we drop a managed table, Hive deletes the data in the table. Whereas in external table, Hive stores meta data about the table, but data is not controlled by Hive; when we drop an internal table, Hived does not delete the data in the table.

4. Conclusions

1.Pro and Cons Native Map-Reduce, Pig and Hive

Map-Reduce

|  |  |
| --- | --- |
| Strengths | Weakness |
| * We can implement complex business logic. * Can be used on both structured and unstructured data. | * Long development type and requires Java. * No straight forward way to achieve join functions. * Need to understand API very well to improve performance. * Has lower level of abstraction. |

Pig

|  |  |
| --- | --- |
| Strengths | Weakness |
| * Development in PIG is much faster, as we might have to deal with fewer Java-level bugs and fewer lines of code. * Higher level of abstraction makes Pig easier to learn and read. | * Unstructured data needs user-defined functions (UDFs) that are written in Java. * Has trouble dealing with complex operations on unbound lists of objects and very flexible schemas |

Hive

|  |  |
| --- | --- |
| Strengths | Weakness |
| * Suitable to perform ad-hoc analysis. * Datasets can be joined easily. * Requires less development time. | * It deals only with structured data. * Has trouble dealing with complex operations |

We have seen how to perform analytics with native MapReduce, Pig and Hive. Based on the type of data and complexity of the business logic, we can choose any of the above discussed methods and tools to accomplish these tasks. One can also use the Hadoop streaming utility that comes with Hadoop distribution. This utility allows you to create and run Map/Reduce jobs with any executable or script as the mapper and/or the reducer. The need to process big data and get insights out of it led to the development of a technology like Hadoop. Tools we have discussed have now reduced the big data challenges to a large extent.

5. References