**Sprint 1 Oracle database as backend and building the dimensional model and implementing Big Data**

**Data Ware house concepts**

General Concepts of data Warehouse

Dimensional modeling

Online Analytical Processing

Data Mining

Best Practices for building data ware house

Project-1

|  |
| --- |
| AARYAN |
| Abhishek Singh |
| Aditya Kumar Das |
| Aditya Pratap Singh |

1. Business Intelligence is playing a vital role for managers to make quality decision to resolve the business needs. BI applications allow managers to query, comprehend, and evaluate existing data to obtain functional knowledge which assist in making improved and informed decisions.

The need of operational data is addressed by the OLTP systems, important for day wise running of business. Decision support queries involve analytics including aggregation, slicing or dicing of data.

Data from OLAP warehouse is extracted and loaded from multiple OLTP data sources.

An entertainment company which specializes in music , films and audio books. This company has multiple offline stores in more than 100 countries. It has few online stores in more than 10 countries. Customers can buy individual products or can subscribe to a package.

Company wants to develop data warehouse and business intelligent system for making better decision using the historical data to support the decision makers and business strategist.

Following changes might be included for designing like:

1. Which is most profitable brand of their business?
2. How do they reuse operational data to support business operations?
3. What are the buying patterns of customers?
4. Which one of the distribution channels is most profitable to the organization?
5. What is the revenue trend of the business process in the current financial year?

Project-2

|  |
| --- |
| Akshat Kumar Sinha |
| Aman Gulwani |
| Aman Jadon |
| Aman Kumar |

1. Business Intelligence is playing a vital role for managers to make quality decision to resolve the business needs. BI applications allow managers to query, comprehend, and evaluate existing data to obtain functional knowledge which assist in making improved and informed decisions.

The need of operational data is addressed by the OLTP systems, important for day wise running of business. Decision support queries involve analytics including aggregation, slicing or dicing of data.

Data from OLAP warehouse is extracted and loaded from multiple OLTP data sources.

An Entertainment company which specializes in video teasers, video songs and movies distributions. It has its business completely on online platform. The size of data which is uploaded on website of company is approx. 72000 hours of new video content every day.

Company wants to develop a system using big data, OLAP tools and data mining to handle queries to handle following issues:

1. Which is most viewed video section of their business?
2. What type of videos are uploaded and viewed by the users?
3. What are the locations where business is doing good?
4. What is the average view duration of advertisements and which category of advertisements are viewed by users?
5. Which videos are skipped by users generally?

Project-3

|  |
| --- |
| Aman Saxena |
| Ankit |
| Anu Sandeep Sathwik Mandavilli |
| Anumeet Saran Singh |
| ARYAMANN SINHA |

1. One of an e-wallet services, that holds credit that can be used to purchase for products on the multiple platforms. Users can receive credit in three different ways:
   1. As a money refund for a product purchase but cancelled now.
   2. Gift card credit as a gift.
   3. So-sorry credit may be provided for poor service experiences.
   4. Gift card and so-sorry credit expires in 6 months , but in 1 year if it is cancellation credit.

**Implementing Big Data**

Hadoop

MapReduce

Hive

Hive Programming

Take any video upload and view site which has a high-volume file uploaded each second. Let assume you have a day partition directory and you need to process around 1 million small files. Use this scenario and discuss how the scenario performs the consolidation on parameter as day.

Perform hive select data from the same partition & insert overwrite in same partition. When Hive re-writes data, it runs a Map-reduce job and reduces the number of files. Draw the Flow work

**Implementing Big Data**

MapReduce

Hive

Hive Programming

Hive external table

Bucketed hive table

Define a hive table form a select query

Load data into a hive table from an HDFS directory

Load a compressed data file into a hive table

Project-4

|  |
| --- |
| Ashwani Sharma |
| Ayush Singh |
| Ayushi Srivastava |
| BHAVESH SINGH JADON |
| DIXIT AGARWAL |

1. Take any video upload and view site which has a high-volume file uploaded each second. Let assume you have a day partition directory and you need to process around 1 million small files. Use this scenario and discuss how the scenario performs the consolidation on parameter as day. Hive selects data from the same partition & insert overwrite in same partition. When Hive re-writes data, it runs a Map-reduce job and reduces the number of files. Now rather than writing the same data select the data from another partition i.e., daily partition. Write this in a temporary partition. On successful writing move this temporary partition data to actual partition with load command. Draw the flow work.

Project-5

|  |
| --- |
| KSHITIZ |
| Purra Gouthami |
| Ravi Choudhary |
| SAHARSH KUMAR |
| SANDHA AKSHARA |

1. Develop a framework for providing solutions to crop disease based on historical data using Hive and Hadoop. Data is to be collected from various sources as laboratory reports, agriculture information pages, and expert recommendation. Irrelevant or the redundant data should be removed. In the further step extract the features from cleaned data, normalization of data is done in order to remove the technical variations. After normalization data is uploaded on HDFS and save in a file that is supported by Hive. Thus, classified data is finally located on the specific place. HiveQL is used to analyze agriculture data based on features and then prioritize the outcome based on crop disease symptoms and in the last a high priority solution is recommended. Draw the flow work for the scenario discussed.

Project-6

|  |
| --- |
| SHRUTI PAREEK |
| SURAJ PRAKASH |
| Thanneeru Laxmi Narayana |
| Tusshar Manish Sinha |
| Teegala Tejaswi |

1. Create a table “user\_info” into any relational database named as “db\_user” like mysql, oracle etc. as follows:

|  |  |
| --- | --- |
| Fieldname | Datatype |
| Id | Bigint |
| Name | Varchar |
| Age | Int |
| Salary | Int |
| Gender | Varchar |
| Mobile | Int |
| Email\_address | Varchar |

Perform the following operations on database “db\_user”

1. List all the databases exist in the relational database using Sqoop
2. Fetch the data from “user\_info” into HDFS using Sqoop
3. Run a query to group all the users in age group of 18-25 and save the result into HDFS directory
4. Create a table into Hive with name “user\_hive”
5. Import “user\_info” into “user\_hive” in Hive
6. Import “user\_info” into Hive creating a new table