How to prepare analytics with large amount of data which is stored in Mysql by using Big Data technologies

Technologies Used:

* Mysql
* Sqoop
* Hadoop HDFS
* Hive
* Pig
* Impala
* Oozie

Work Flow:

By using sqoop import Dump the mysql data into HDFS,loading the data from HDFS to Either Hive or Pig then run the query scripts and get the result and dump the result back into agian Mysql by using Sqoop export and automate the entire process by using oozie workflow management





Mysql:

Our mysql sales database having three types of tables related to sales and they are Product,Inventory,Sale and there schemas are,

Product Table Schema:

+--------------+--------------+------+-----+---------+----------------+

| Field | Type | Null | Key | Default | Extra |

+--------------+--------------+------+-----+---------+----------------+

| id | bigint(20) | NO | PRI | NULL | auto\_increment |

| version | bigint(20) | NO | | NULL | |

| brand\_name | varchar(255) | NO | | NULL | |

| category | varchar(255) | NO | | NULL | |

| price | int(11) | NO | | NULL | |

| product\_name | varchar(255) | NO | | NULL | |

| weight | varchar(255) | NO | | NULL | |

+--------------+--------------+------+-----+---------+----------------+

Sale Table Schema:

+----------------+--------------+------+-----+---------+----------------+

| Field | Type | Null | Key | Default | Extra |

+----------------+--------------+------+-----+---------+----------------+

| id | bigint(20) | NO | PRI | NULL| auto\_increment |

| version | bigint(20) | NO | | NULL | |

| brand\_name | varchar(255) | NO | | NULL | |

| item\_purchased | bigint(20) | NO | | NULL | |

| product\_id | bigint(20) | NO | | NULL | |

| purchase\_date | datetime | NO | | NULL | |

+----------------+--------------+------+-----+---------+----------------+

Inventory Table Schema:

+-----------------+------------+------+-----+---------+----------------+

| Field | Type | Null | Key | Default | Extra |

+-----------------+------------+------+-----+---------+----------------+

| id | bigint(20) | NO | PRI | NULL| auto\_increment |

| version | bigint(20) | NO | | NULL | |

| inventory | bigint(20) | NO | | NULL | |

| product\_id | bigint(20) | NO | | NULL | |

| production\_date | datetime | NO | | NULL | |

+-----------------+------------+------+-----+---------+----------------+

Sqoop:

By using sqoop tool we are going to transfer the data from mysql database to HDFS and vise versa.You can pass directly the mysql data into either Hive,Pig also.For Sqoop import and export commands please check the sqoop script file

After dumping the data into Hdfs Loading the data into Either Pig or Hive by using Hive and Pig script files given here

Hive create the tables and load the data into tables by using **Table\_Schema.sql** file

Run the corresponding script files for

1) Total Sales by Year

**year\_sales.pig**

**sales\_by\_year.sql**

2) Total Sales by Brand and Year

**brand\_sales\_by\_year.pig**

**sales\_by\_brand.sql**

3) Top 5 sales

**top\_5\_sales.pig**

**top\_5\_ sales.sql**

4) Monthly Sales

**month\_sales.pig**

**month\_sales.sql**

5) Inventory Left

**product\_brand\_inventory\_left.pig**

**inventry\_left.sql**

Once getting the results dump them back into mysql by using Sqoop export command

**Note**: Before dumping the data into mysql you need the create the table based on your output records otherwise it throws error

To make the things automate use oozie workflow it will controls the total workflow, for creating oozie workflow follow the workflow.xml for pig and workflow.xml for hive in oozie folder and it’s instructions