Hadoop

Technical Skill Set: Programming Languages Apache Hadoop, Python, shell scripting, SQL Technologies Hive, Pig, Sqoop, Flume, Oozie, Impala, hdfs Tools Dataiku, Unravel, Cloudera, Putty, HUE, Cloudera Manager, Eclipse, Resource Manager Initial Learning Program: Tata Consultancy Services: June 2015 to August 2015 Description: This is a learning program conducted by TCS for the newly joined employees, to accomplish them to learn the working standard of the organization. During this period employee are groomed with various technical as well as ethical aspects. Education Details

B.E. Electronics & Communication Indore, Madhya Pradesh Medi-caps Institute of Technology & Management

Hadoop developer

hadoop,hive,sqoop,flume,pig,mapreduce,python,impala,spark,scala,sql,unix.

Skill Details

APACHE HADOOP SQOOP- Experience - 31 months

Hadoop- Experience - 31 months

HADOOP- Experience - 31 months

Hive- Experience - 31 months

SQOOP- Experience - 31 months

python- Experience - Less than 1 year months

hdfs- Experience - Less than 1 year months

unix- Experience - Less than 1 year months

impala- Experience - Less than 1 year months

pig- Experience - Less than 1 year months

unravel- Experience - Less than 1 year months

mapreduce- Experience - Less than 1 year months

dataiku- Experience - Less than 1 year monthsCompany Details

company - Tata Consultancy Services

description - Project Description

Data warehouse division has multiple products for injecting, storing, analysing and presenting data. The Data Lake program is started to provide multi-talent, secure data hub to store application's data on Hadoop platform with strong data governance, lineage, auditing and monitoring capabilities. The object of the project is to provide necessary engineering support to analytics and application teams so that they can focus on the business logic development. In this project, the major task is to set up the Hadoop cluster and govern all the activities which are required for the smooth functioning of various Hadoop ecosystems. As the day and day data increasing so to provide stability to the ecosystem and smooth working of it, Developing and automating the various requirement specific utilities.

Responsibility 1. Developed proactive Health Check utility for Data Lake. The utility proactively checks the smooth functioning of all Hadoop components on the cluster and sends the result to email in HTML format. The utility is being used for daily Health Checks as well as after upgrades.

2. Getting the data in different formats and processing the data in Hadoop ecosystem after filtering the data using the appropriate techniques.

3. Developed data pipeline utility to ingest data from RDBMS database to Hive external tables using Sqoop commands. The utility also offers the data quality check like row count validation.

4. Developed and automated various cluster health check, usage, capacity related reports using Unix shell scripting.

5. Optimization of hive queries in order to increase the performance and minimize the Hadoop resource utilizations.

6. Creating flume agents to process the data to Hadoop ecosystem side.

7. Performed benchmark testing on the Hive Queries and impala queries.

8. Involved in setting up the cluster and its components like edge node and HA implementation of the services: Hive Server2, Impala, and HDFS.

9. Filtering the required data from available data using different technologies like pig, regex Serde etc.

10. Dataiku benchmark testing on top of impala and hive in compare to Greenplum database.

11. Moving the data from Greenplum database to Hadoop side with help of Sqoop pipeline, process the data to Hadoop side and storing the data into hive tables to do the performance testing.

12. Dealing with the Hadoop ecosystem related issues in order to provide stability to WM Hadoop ecosystem.

13. Rescheduling of job from autosys job hosting to TWS job hosting for better performance.

Declaration:

I herweby declare that the above mentioned information is authentic to the best of my knowledge

company - Tata Consultancy Services

description - Clients: 1. Barclays 2. Union bank of California (UBC) 3. Morgan Stanley (MS)

KEY PROJECTS HANDLED

Project Name ABSA- Reconciliations, UBC and WMDATALAKE COE

company - Tata Consultancy Services

description - Project Description

Migration of data from RDBMS database to Hive (Hadoop ecosystem) . Hadoop platform ability with strong data governance, lineage, auditing and monitoring capabilities. The objective of this project was to speed up the data processing so that the analysis and decision making become easy. Due to RDBMS limitations to process waste amount of data at once and produce the results at the earliest, Client wanted to move the data to Hadoop ecosystem so that they can over-come from those limitations and focus on business improvement only.

Responsibility 1. Optimising the SQL queries for those data which were not required to move from RDBMS to any other platform.

2. Writing the Hive queries and logic to move the data from RDBMS to Hadoop ecosystem.

3. Writing the hive queries to analyse the required data as per the business requirements.

4. Optimization of hive queries in order to increase the performance and minimize the Hadoop resource utilizations.

5. Writing the sqoop commands and scripts to move the data from RDBMS to Hadoop side.

company - Tata Consultancy Services

description - Project Description

Create recs and migrating static setup of reconciliations from 8.1 version to 9.1 version of the environment Intellimatch.

Responsibility 1. Have worked on extracting business requirements, analyzing and implementing them in developing Recs 2. Worked on migrating static setup of reconciliations from 8.1 version to 9.1 version of the environment Intellimatch.

3. Done the back end work where most of the things were related to writing the sql queries and provide the data for the new recs.

Project Name   PSO