PROFESSIONAL SUMMARY

- Around 6 years of overall experience, in software design, development, maintenance, testing, and troubleshooting of enterprise applications.
- Over 4 plus years of experience in design, development, and maintenance and support of Big Data Analytics using Hadoop Ecosystem components like HDFS, Hive, Pig, Sqoop, Zookeeper, Map Reduce, and Oozie.
- Strong working experience with ingestion, storage, processing, and analysis of big data.
- Good Experience in writing Map Reduce programs using Java.
- Expertise in writing Hadoop Jobs for analyzing data using Hive and Pig.
- Good Experience in designing the Nifi flows for data routing, transformations, and mediation logic.
- Successfully loaded files to HDFS from Oracle, SQL Server and Teradata using SQOOP.
- Experience with SQL, PL/SQL, and database concepts
- Good Experience with job workflow scheduling like Oozie and Airflow.
- Good understanding of NoSQL databases.
- Experience on creating databases, tables and views in HIVE, Impala and Snowflake.
- Experience with performance tuning on map reduce and hive jobs.
- Load and transform large sets of structured, semi-structured and unstructured data using Hadoop ecosystem components.
- Worked with Sqoop in importing and exporting data from different databases like MySQL, Oracle into HDFS and vice versa.
- Experience in working with different data sources like Flat files, XML, JSON files and Databases.
- Experience in database design, entity relationships, database analysis, programming SQL, stored procedure's PL/ SQL packages, and triggers in Oracle.
- Experience in various phases of Software Development Life Cycle (Analysis, Requirements gathering, Designing) with expertise in documenting various requirement specifications, functional specifications, Test Plans, Source to Target mappings, SQL Joins.

EXPERIENCE

Oct 2018 – Till Date

Client: Nike
Role : Senior Data Engineer

The main goal of the project is to migrate the existing data from Oracle to AWS S3 and perform ETL operations and store the data in the Snowflake that uses the SQL database engine designed for the cloud and also help the BI users to gain the business insights from it for preparing the dashboards .

Responsibilities:

• Evaluate, extract/transform data for analytical purpose within the context of Big data environment.
• Migrating the existing data from Oracle to AWS and perform ETL operations on it using Qubole.
• Responsible for using Hadoop and spark for data warehouse applications to maintain large datasets in AWS S3 and decide on engineering tools based on recommendations.
• Design and develop spark scripts to gather data insights as per business requirements and collaborate with other teams on integration needs/design.
• Facilitate or perform application support, problem solving, and issue resolution with internal and external resources.
• Resolve big data issues and determine options for issue resolution and risk mitigation.
• Working with Avro and Parquet files formats and used various compression techniques to leverage the storage in HDFS.
• Worked with the Data Scientist team to build pipelines for their Machine Learning models.

SKILLS:

Hadoop Ecosystem:

- MapReduce
- Hive
- Pig
- Flume
- Sqoop
- Oozie

Monitoring and Automation:

- Nagios
- Ganglia
- Cloudera Manager
- Autosys
- Airflow

Databases:

- Oracle 9i/10g/11g
- SQL Server 2005/2008

Languages:

- Python
- C
- Java

Reporting Tools:

- Framework manager
- Tableau

• Review and approve performance test results, recommendations, and tuning results. Oversee and is responsible for the creation of test plans, test execution, and validation of test results.
• Responsible for EMR Cluster creation, administration, sizing and configuration.
• Created Spark jobs to see trends in data usage by users.
• Worked with SCRUM team in delivering agreed user stories on time for every sprint.
• Development and unit testing on Hadoop and AWS ecosystem.
• Automate and monitor the ETL process and applications.
• Good Knowledge on Spark framework on both batch and real time data processing.
• Designed, developed ETL workflow and automated using Autosys.

Environment: Qubole, AWS, Snowflake, Spark, Airflow, Databricks, CICD.

Employer : Data Capital Inc. **March 2018 - Oct 2018**
Client: Walmart
Role : Big Data Developer

The main goal of the project is to migrate the existing data from Teradata/ Mainframes /Oracle to Hadoop and perform ETL operations that helps Walmart Business with the key insights and faster decisions using the cutting edge visualization tools like ThoughtSpot which is used as BI tool for holding the latest data to drill down to minute grain level.

Responsibilities:

- Migrating the existing data from Mainframes/Teradata/Oracle to Hadoop and perform ETL operations on it.
- Designed and Implemented Sqoop incremental imports, delta imports on tables without primary keys and dates from Teradata and appends directly into Hive Warehouse.
- Used aorta connector to load History data into Teradata WM3/WMG boxes.
- Experience in loading data from different sources into HDFS using internal aorta application.
- Worked on POC to evaluate the performance of multi tenancy tables vs standalone tables and performance of views on top of these.
- Used Automic workflow engine to manage interdependent Hadoop jobs and to automate several types of Hadoop jobs such as map-reduce Hive, Sqoop and Spark jobs.
- Working with Avro and Parquet file formats and used various compression techniques to leverage the storage in HDFS.
- Using the Mainframe SerDe's and Avro SerDe's for serialization and de-serialization in hive to parse the contents
- Designed and developed ETL workflow using Automic for scheduling.

Environment: Hortonworks, Autosys, Automic, Oozie, Mainframes, Teradata, Oracle.

Employer : Cloudwick Technologies, Newark, California **Nov 2017 - March 2018**
Role : Bigdata / Spark developer

The main objective of the project is to perform analytics and gain insights from the data which is being moved from Teradata and Netezza to AWS Cloud environment. Our responsibility is to build an ETL data pipeline to load it into dataware house platforms like Redshift and other sql based databases (Hive and Presto) to perform analytics on the data.

Responsibilities:

NOSQL database:

- Cloudant
- Hbase

Other Tools:

- SQL Management Studio
- Eclipse, Serena Version Control Tool

- Used Pyspark to read the data from S3 and perform various transformations to prepare the data for loading.
- Develop python scripts for Data Quality/Standardization checks.
- Worked with Spark for various transformations.
- Experience in designing and developing applications in Spark using python to compare the performance of Spark with Hive and SQL/Oracle.
- Load the data from S3 to Hive and presto using different file formats like JSON and ORC
- Load the data into redshift using Pyspark for generating the quarterly performance reports.
- To facilitate the BI team to generate reports using Tableau /SAS based on the data present in Redshift.
- Designed and developed jobs to validate the data post migration such as reporting fields from source and designation systems using Spark SQL RDDs and Data Frames/Datasets.
- Used Spark SQL on data frames to access hive tables into spark for faster processing of data.

Environment: AWS, Oracle, Pyspark, Redshift, Tableau, Presto, Hive

---

Employer : Sparsity Systems LLC. **Jan 2017 - Oct 2017**
Client: Universal Orlando
Role : Hadoop Engineer

The main objective of the project is store the hotel and ticketing information in Hadoop and perform various analytics for business and we also keep track of the visitors who visit more often on monthly and annually based. With the help of the ticketing data, we analyze the visitors who are coming into the park more frequently and give them various offers on the tickets and universal hotels information and also provide them the wait times for each of the rides in the theme park, the BI users create dashboards with the help of Daily ticket sales.

Responsibilities:

- Used Nifi to ingest the data from various sources into the datalake.
- Worked with Spark for various transformations.
- Created Hive managed and external tables.
- Used Kafka for streaming application.
- Created topics in kafka broker which gets the data from sources with the help of Nifi and Spark job consumes it and pushes it into IBM Cloudant Database.
- Worked with Partitioning, bucketing and other optimizations in hive.
- Worked with ORC, JSON file formats and used various compression techniques to leverage the storage in HDFS.
- Developed and implemented core API services using Spark with Scala.
- Used Rally to keep the track of the user stories and tasks for completing in each sprint.
- Worked on ingesting the data from hive to spark and create data frames in spark then updating it into IBM Cloudant Database.
- Used Pydev to perform business logic environment to call the REST API’s to update/create the documents in the IBM Cloudant Database
- Also prepared the data with the help of Paxata (a data preparation tool) for our Business users.
- Worked on various production issues during the month end support and provide the resolutions without missing any SLA.
- Used GitHub to set the overall direction of the project and track the progress of the project.
- Used Paxata for delivering the data to the BI users for creating the dashboards for the Daily Sales ticket of the theme park.

Environment: Hortonworks, Nifi, Kafka, IBM Cloudant, Paxata, GitHub.

Employer : Sparsity Systems LLC. **March 2016 – Jan 2017**
Client: NBCUniversal Media LLC
Role : Hadoop /Spark Developer

The main objective of the project is to perform analytics and gain insights from the data which is being moved from Teradata and Netezza to AWS Cloud environment. Our responsibility is to build an ETL data pipeline to load it into data warehouse platforms like Redshift and other SQL based databases (Hive and Presto) to perform analytics on the data.

Responsibilities:

- Used Pyspark to read the data from S3 and perform various transformations to prepare the data for loading.
- Develop python scripts for Data Quality/Standardization checks.
- Worked with Spark for various transformations.
- Experience in designing and developing applications in Spark using python to compare the performance of Spark with Hive and SQL/Oracle.
- Load the data from S3 to Hive and presto using different file formats like JSON and ORC
- Load the data into redshift using Pyspark for generating the quarterly performance reports.
- To facilitate the BI team to generate reports using Tableau /SAS based on the data present in Redshift.
- Designed and developed jobs to validate the data post migration such as reporting fields from source and designation systems using Spark SQL RDDs and Data Frames/Datasets.
- Used Spark SQL on data frames to access hive tables into spark for faster processing of data.

Environment: Alteryx, Bedrock, Hortonworks, Paxata, AWS S3, Hive. Teradata.

EDUCATIONAL QUALIFICATION

Master of Science, Computer Science Engineering
GA at University of Michigan, Flint **(Aug 2014 – Dec 2015)**

Academic Projects:

- **Software Engineering:** Designed an “Integrated Home Entertainment System” which helps the user to play games and watch his favorite movies.
- **Data warehousing:** Extracted relevant data using JSON and loaded it into the warehouse and implemented the OLAP and mining functions.
- **Databases**: Designed “Coaching Management System” website, developed using Java as the programming language and MYSQL as the database.

Bachelors in Computer Sciences and Engineering, Jawaharlal Nehru Technological University, India
(2010- 2014)