Hadoop Big Data Engineer and Developer with skills in legacy Hadoop Ecosystems, Cloudera Hadoop, Hortonworks Hadoop and Amazon Web Services. Skilled in use of Spark, Spark Streaming, Spark SQL, Kafka, and Kibana.

5 Years Experience in Big Data and I.T.

Professional Summary

* 5 years’ experience in Hadoop Big Data and 5 years’ experience in I.T.
* Seasoned experience in project management (Agile/Scrum, Waterfall and various Agile processes).
* Proven success in team leadership, focusing on mentoring team members, and managing task for efficiency.
* Worked with various stakeholders for gathering requirements to create as-is and as-was dashboards.
* Modified existing and added new functionalities to Financial and Strategic summary dashboards.
* Analyzed the MS-SQL data model and provided inputs for converting the existing dashboards that used Excel as a data source.
* Created dashboards for TNS Value manager in Tableau using various features of Tableau like Custom-SQL, Multiple Tables, Blending, Extracts, Parameters, Filters, Calculations, Context Filters, Data source filters, Hierarchies, Filter Actions, Maps etc.
* Involved in Performance tuning the data heavy dashboards and reports for optimization using various options like Extracts, Context filters, writing efficient calculations, Data source filters, Indexing and Partitioning in data source etc.
* Writing SQL queries for data validation of the reports and dashboards as necessary.
* Recommended and used various best practices to improve dashboard performance for Tableau server users.
* Worked with Data Lakes and Big Data ecosystems (Hadoop, Spark, Hortonworks, Cloudera)
* Expert with BI tools like Tableau and PowerBI, data interpretation, modeling, data analysis, and reporting with the ability to assist in directing planning based on insights.
* Track record of results as a project manager in an Agile methodology using data-driven analytics.
* Used to working in a production environment, managing migrations, installations, and development.
* Have managed teams ranging from 5 to 20 members with on-site and remote members, across multiple time zones, in a culturally diverse environment.
* Expert with design of custom reports using data extraction and reporting tools, and development of algorithms based on business cases.
* Able to design new custom solutions to solve business issues and advance goals.
* Keeps managers and stakeholders apprised of project status and informed of any pertinent events.
* Expert in project management tools such as Microsoft Team Foundation Server (TFS), Jira, and MS Project.
* Knowledgeable of database technologies and frameworks involving structured data, unstructured data, and semi-structured data as well as various storage formats such as RDMS and data lakes.

Technical Skills

|  |  |
| --- | --- |
| IDE: | Jupyter Notebooks (formerly iPython Notebooks), Eclipse, IntelliJ, PyCharm |
| PROJECT METHODS: | Agile, Kanban, Scrum, DevOps, Continuous Integration, Test-Driven Development, Unit Testing, Functional Testing, Design Thinking, lean, Six Sigma |
| HADOOP DISTRIBUTIONS: | Hadoop,Cloudera Hadoop, Hortonworks Hadoop |
| CLOUD PLATFORMS: | Amazon AWS - EC2, SQS, S3, MapR, Elastic Cloud |
| CLOUD SERVICES: | Solr Cloud, Databricks, Datastax |
| CLOUD DATABASE & TOOLS: | Redshift, DynamoDB, Cassandra, Apache Hbase, SQL |
| PROGRAMMING LANGUAGES: | Spark, Spark Streaming, Java, Python, Scala, PySpark, PyTorch |
| SCRIPTING: | Hive, Pig, MapReduce, SQL, Spark SQL, Shell Scripting |
| CONTINUOUS INTEGRATION (CI CD): | Jenkins |
| VERSIONING: | Git, GitHub |
| PROGRAMMING METHODOLOGIES: | Object-Oriented Programming, Functional Programming |
| FILE FORMAT AND COMPRESSION: | CSV, JSON, Avro, Parquet, ORC |
| FILE SYSTEMS: | HDFS |
| ETL TOOLS: | Apache Camel, Flume, Kafka, Talend, Pentaho, Sqoop |
| DATA VIZIUALIZATION TOOLS: | Tableau, Kibana |
| SEARCH TOOLS: | Apache Lucene, Elasticsearch |
| SECURITY: | Kerberos, Ranger, Blockchain |
| AWS: | AWS Lambda, AWS S3, AWS RDS, AWS EMR, AWS Redshift, AWS S3, AWS Lambda, AWS Kinesis, AWS ELK, AWS Cloud Formation, AWS IAM |
| Data Query: | Spark SQL, Data Frames |

Professional Experience

**Big Data Engineer**

**USAA, San Antonio, Texas September 2019- March 2020**

* Performance tuning of Spark jobs for setting batch interval time, level of parallelism, and memory tuning.
* Developed POC using Scala & deployed on Yarn cluster, compared the performance of Spark, with Hive and SQL.
* Hive for queries and incremental imports with Spark and Spark jobs for data processing and analytics.
* Install and configure Kafka cluster and monitoring the cluster; Architected a light weight Kafka broker; integration of Kafka with Spark for real time data processing.
* Build a Spark proof of concept with Python using PySpark
* Spark applications using Spark Core, Spark SQL and Spark Streaming API
* Extracted the needed data from the server into Hadoop file system (HDFS) and bulk loaded the cleaned data into HBase using Spark.
* Built a prototype for real-time analysis using Spark streaming and Kafka.
* Worked on Spark SQL to check the data; Wrote Spark applications for data validation, cleansing, transformation, and custom aggregation.
* Hands-on experience with Spark Core, Spark SQL and Data Frames/Data Sets/RDD API.
* Spark jobs, Spark SQL and Data Frames API to load structured data into Spark clusters.
* Created a Kafka broker which uses schema to fetch structured data in structured streaming.
* Spark streaming to receive real time data using Kafka.
* Interacted with data residing in HDFS using Spark to process the data.
* Participated in various phases of data processing (collecting, aggregating, moving from various sources) using Apache Spark.
* Handled structured data via Spark SQL then stored into Hive tables for downstream consumption.
* Accessed Hadoop file system (HDFS) using Spark and managed data in Hadoop data lakes with Spark.
* Set-up Jenkins CI server for use by developers for continuous integration.
* Handled structured data with Spark SQL to process in real time from Spark Structured Streaming.
* Used Apache Spark framework with Scala mainly.
* Support for the clusters, topics on the Kafka manager.
* Integrated Kafka with Spark Streaming for real time data processing

**AWS Big Data Engineer**

**NCR Corporation, Atlanta, GA June 2017- September 2019**

* Configured access for inbound and outbound traffic RDS DB services, DynamoDB tables, EBS volumes to set alarms for notifications or automated actions on AWS.
* Implemented AWS IAM user roles and policies to authenticate and control access.
* Specified nodes and performed the data analysis queries on Amazon redshift clusters on AWS.
* Developed AWS Cloud Formation templates to create custom infrastructure of our pipeline.
* Working on AWS Kinesis for processing huge amounts of real time data.
* Developed multiple Spark Streaming and batch Spark jobs using Scala and Python on AWS.
* Created multiple batch Spark jobs using Java.
* Ingestion data through AWS Kinesis Data Stream and Firehose from various sources to S3.
* RDS, Cloud Formation, AWS IAM and Security Group in Public and Private Subnets in VPC.
* Worked with AWS Lambda functions for event-driven processing to various AWS resources.
* Automated AWS components like EC2 instances, Security groups, ELB, RDS, Lambda and IAM through AWS cloud Formation templates.
* Processed multiple terabytes of data stored in AWS using Elastic Map Reduce (EMR) to AWS Redshift.
* Implemented security measures AWS provides, employing key concepts of AWS Identity and Access Management (IAM).
* Installed, Configured and Managed AWS Tools such as ELK, Cloud Watch for Resource Monitoring.
* AWS EMR to process big data across Hadoop clusters of virtual servers on Amazon Simple Storage Service (S3).
* Launched and configured The Amazon EC2 (AWS) Cloud Servers using AMI's (Linux/Ubuntu) and configuring the servers for specified applications.
* Responsible for Designing Logical and Physical data modelling for various data sources on AWS Redshift.

**Big Data Developer**

**Capco, New York, NY January 2015-June 2017**

* Wrote incremental imports into Hive tables.
* Created Hive tables to store the processed results in a tabular format.
* Installed Oozie workflow engine to run multiple Hive Jobs.
* Wrote Hive queries and optimized the Hive queries with Hive QL.
* ETL to Hadoop file system (HDFS) and wrote HIVE UDFs.
* Experienced in importing real-time logs to HDFS using Flume.
* Created UNIX shell scripts to automate the build process, and to perform regular jobs like file transfers.
* Cloudera Manager used for installation of Cloudera Cluster and performance monitoring.
* Made incremental imports to Hive with Sqoop.
* Managed Hadoop clusters and check the status of clusters using Ambari.
* Moved Relational Database data using Sqoop into Hive Dynamic partition tables using staging tables.
* Initiated data migration from/to traditional RDBMS with Apache Sqoop.
* Developed scripts to automate the workflow processes and generate reports.
* Transfered data between a Hadoop ecosystem and structured data storage in a RDBMS such as MySQL using Sqoop.
* Wrote Hive queries and wrote custom UDF’s.
* Writing Hive Queries for analyzing data in Hive warehouse using Hive Query Language.
* Configured Hadoop components (HDFS, Zookeeper) to coordinate the servers in clusters.
* Hive partitioning, bucketing, and joins on Hive tables, utilizing Hive SerDe’s.

**Big Data Developer**

**ESL Federal Credit Union, Rochester, NY November 2013-January 2015**

* Wrote shell scripts to automate workflows to pull data from various databases into Hadoop.
* Loaded into Hbase tables and Hive tables consumption purposes.
* Performed upgrades, patches and bug fixes in Hadoop in a cluster environment.
* Wrote Hive queries and wrote custom UDF’s.
* Writing Hive Queries for analyzing data in Hive warehouse using Hive Query Language.
* Interaction with NOC team to work with Hadoop to provide large-scale solutions.
* Used Cloudera Manager for installation and management of single-node and multi-node Hadoop cluster.
* Extracted metadata from Hive tables with Hive QL.
* Built the Hive views on top of the source data tables, and built a secured provisioning framework for users to access the data through Hive based views.
* Wrote shell scripts for automating the process of data loading.
* Experience collecting log data from various sources and integrating it into HDFS using Flume; staging data in HDFS for further analysis.
* Wrote the Hive scripts to process the HDFS data.
* Created Hive queries to spot emerging trends by comparing Hadoop data with historical metrics.
* Installed and configured Tableau Desktop to connect to the Hortonworks Hive Framework (Database) which contains the Bandwidth data
* Created Hive tables, loading with data and writing Hive queries.

Education & Training

Bachelor’s Degree in Management Information System with Computer Science minor – University of Georgia

Certifications

* Lean Six Sigma Yellow Belt

----Pending