Professional Profile

**7 years programming with 5 years in Hadoop/Big Data**

* Understanding of distributed systems, HDFS architecture, internal working details of MapReduce and Spark processing frameworks.
* Understanding of Hadoop big data architectures, data movement technologies, database partitioning, database optimization, and building communication channels between structured and unstructured databases.
* Understanding of big data concepts and use of cloud technologies and tools.
* Pandas experience with operational monitoring of clusters and applications.
* Setup and Config Data Lake on Bigdata platforms like Cloudera, Hortonworks, and MapR.
* Skill using Open source software like Spark, Flume, and Kafka.
* Proficient performing importing/exporting actions between SQL (Oracle, MySQL) / NoSQL (Realm, MongoDB) databases and HDFS using Sqoop.
* Proficient with building tools like Apache Maven and SBT.
* Hands-on expertise in Hadoop components - HDFS, MapReduce, Hive, Impala, Flume, Sqoop and HBase.
* Implemented, set-up and worked on various Hadoop Distributions (Cloudera, Hortonworks, Amazon AWS).
* Knowledgeable of deploying the application jar files into AWS instances.
* Knowledgeable of Hadoop Architecture and Hadoop components (HDFS, MapReduce, NameNode, DataNode, ResourceManager, NodeManager.
* Knowledgeable of installation and configuration of Hadoop, Hive, on AWS EC2.
* Installed and configured of Hive, HBase, Flume and MongoDB, Cassandra and Kafka on Hadoop clusters.

Professional Skills

|  |  |
| --- | --- |
| **Programming**  Java, SQL, Scala, Python, Java, C/C++, SQL  **Database**  Apache Cassandra, AWS DynamoDB, MongoDB, AWS AuroraDB, AWS Redshift, AWS RDS, SQL, MySQL, NoSQL, Oracle.  **Open Source Distributions**  AWS, Kali, Cisco  **Amazon Stack**  AWS, EMR, EC2, EC3, SQS, S3, DynamoDB, Redshift, Cloud Formation  **Systems**  Windows Active Directory, Windows Server 2003, 2008, 2008R2, 2012/2012R2, Red Hat Linux 6-7, IBM AIX, HP, CentOS, Ubuntu  **Virtualization**  VMWare, vSphere, Virtual Machine/ Big Data VM, VirtualBox, Oracle Big Data Lite Virtual Machine v 4.9  **Data Pipelines/ETL**  Flume, Apache Kafka, Atom, Heka, Logstash, Stitch, Talend, Ketl, Pentaho Data Integration (Kettle), Jaspersoft, CloverETL | **Distributions**  Cloudera, Hortonworks. AWS, MapR  **Hadoop Ecosystem**  Hadoop, Hive, Spark, Maven, Ant, Kafka, HBase, yarn, Flume, Zookeeper, Impala. HDFS, Pig, Oozie, Tez, Zookeeper, Apache Airflows  **Search Tools**  Apache Solr/Lucene, Elasticsearch/ Kibana  **File Formats**  Parquet, Avro  **File Compression**  Snappy, Gzip, ORC  **Data Mining**  RapidMiner, IBM SPSS Modeler, Oracle Data Mining  **Data Cleansing**  DataCleaner, Winpure Data Cleaning Tool, Patnab, OpenRefine, Drake  **Software**  Nessus, SET, API, Metasploit, WIreShark, VMWare, vSphere, AppDynamics, Confluence, Jira, RabbitMQ, Minsoft, Nagios, Cloudclock |

Professional Experience

**SENIOR BIG DATA ENGINEER March 2017 - Present**

**MORONGO CASINO Cabazon, CA**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

* PySpark stop words removal implementation
* Developed Spark streaming and batch jobs in Scala and Python on AWS.
* Developed udfs, data frames, and rdds in PySpark
* Tuning and operating Spark and related technologies
* Map Reduce algorithm in PySpark
* Exported analyzed data to RDBMS with Sqoop for visualization/reporting
* Migrated data from RDBMS into Hadoop cluster using Hive, Pig, Flume and Sqoop
* Transformed data model into and implemented database design
* Used Hive to simulate data warehouse for performing client-based transit system analytics.
* Built continuous Spark streaming ETL pipeline with Spark, Kafka, and Scala,
* Developed SQL queries to Insert, Update and Delete data in Database.
* Worked on large data warehouse Analysis services servers and developed the different reports for the analysis from that servers.
* Worked on stored procedures, triggers and on substantial number of business analytical functions.
* ETL data using Sqoop between RDBMS and Hadoop
* Fetching the live stream data to write to HBase table using Spark Streaming and Apache Kafka.
* Involved in complete Big Data flow of the application starting from data ingestion from upstream to HDFS, processing the data into HDFS using Spark Streaming.

**DATA ENGINEER May 2016 - March 2017**

**APPLE Sunnyvale, CA**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

* Developed Spark code using Scala and Spark-SQL/Streaming for faster processing of data.
* Created modules for Spark streaming in data into Data Lake using Storm and Spark.
* Transferred Streaming data from different data sources into HDFS and HBase using Apache Flume.
* Fetching the live stream data from DB2 to HBase table using Spark Streaming and Apache Kafka.
* Used a proprietary Apple code repository and version control.
* Managed commits to Apple repository to ensure quality
* Involved in complete Big Data flow of the application starting from data ingestion from upstream to HDFS, processing the data into HDFS using Spark Streaming.
* Implemented data ingestion and cluster handling in real time processing using Kafka.
* Integrated Kafka with Spark Streaming for real time data processing
* Moved transformed data to Spark cluster where the data is set to go live on the application using Kafka.
* Created a Kafka producer to connect to different external sources and bring the data to a Kafka broker.
* Java language development for web-based applications running on Linux.
* Developed Spark code using Scala and Spark-SQL/Streaming for faster processing of data.
* Configured Spark Streaming to receive real time data and store the stream data to HDFS.
* Extract Real time feed using Kafka and Spark Streaming and convert it to RDD and process data in the form of Data Frame and save the data as Parquet format in HDFS.
* Used different file formats like text files, sequence files, and Avro.

**BIG DATA ENGINEER January 2015 – May 2016**

**ITRANSITION Austin, TX**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

* Implement Hive queries in Spark-SQL integrating Spark environment in Scala.
* Created hive external tables to store pig output
* Import and export data using Sqoop from HDFS to AWS, EMR, EC2, data pipeline, redshift, AWS cli
* Created PySpark code to implement a map-reduce algorithm
* Archive data to Hadoop cluster allowing searching
* Serialize/deserialize data for Kafka producer and consumer using Avro algorithm
* Apache Hadoop core, flume, yarn, hive, Kafka, and maven
* Cassandra SQL, pl/sql to creating tables, views, indexes, stored procedures, etc
* Query Hive with Spark-SQL int integrated Spark environment using Scala
* Design ETL workflows using Python and Scala for data processing in HDFS & MongoDB
* Performance optimization of Hive queries
* Converted Hive/SQL queries into Spark transformations using Spark RDDs, Python and Scala.
* Developed the build script for the project using SBT Framework.

**DATA ENGINEER February 2013 – September 2014**

**GLOBAL STRATEGY GROUP** **Seattle, WA**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

* PySpark stop words removal implementation
* Developed mapreduce routines in Java for data transformations
* Imported unstructured data into HDFS using Spark
* Creation of Spark udf functions in Python and Scala
* Troubleshoot and tune sql, Python, Scala, PIG, Hive, etc.
* Used Scala to collect data from web api
* Bash script with awk formatted text to send metrics to influxdb.
* Wrote Sqoop scripts to inbound and outbound data to HDFS and validated the data before loading to check the duplicated data.
* Imported data using Sqoop to load data from MySQL to HDFS on regular basis.
* Involved in transforming data from legacy tables to HDFS, and HBase tables using Sqoop.
* Connected various data centers and transferred data between them using Sqoop and various ETL tools.
* Extracted the data from RDBMS (Oracle, MySQL) to HDFS using Sqoop.
* Involved in transforming the relational database to legacy labels to HDFS, and HBASE tables using Sqoop and vice versa.
* Developed and ran Map-Reduce jobs on YARN and Hadoop clusters to produce daily and monthly reports per requirements.
* Developed MapReduce jobs using Java for data transformations.

Education

**San Diego State University; San Diego, CA**

Bachelor’s of Science in Computer Science