MIYA MA | HADOOP | 628-222-5612| weima3389@gmail.com

MIYA MA

SENIOR BIG DATA ENGINEER

PROFESSIONAL PROFILE

CONTENTS

5 years of experience in Big Data

- Developed Cloud-based Big Data Architecture using Hadoop and AWS which created the foundation of this Enterprise Analytics initiative in a Hadoop-based Data Lake.
- Created variation of the lambda architecture consisting of near real-time using Spark SQL.
- Apache Open source version with Mesos job scheduler.
- Created multi-node Hadoop and Spark clusters in AWS instances to generate terabytes of data and stored it in AWS HDFS.
- Deployed the application jar files into AWS instances.
- Developed a task execution framework on EC2 instances using SQL and DynamoDB.
- Captured and transformed real-time data from Amazon Aurora into a suitable format for Scalable analytics.
- Investigation of machine learning at scale using Amazon SageMaker on AWS.
- Used Cloud formation scripting to automate resource creation.
- worked with Amazon EC2, Amazon S3, Amazon RDS, Amazon Elastic Load Balancing, Amazon SQS, and other services of the AWS family.
- 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.
- Worked with Data Lakes and Big Data ecosystems (Hadoop, Spark, Hortonworks, Cloudera)
- Experienced in 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.
- Able to design new custom solutions to solve business issues and advance goals.
- 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.
- Manipulated and analysed complex, high volume, and high dimensional data in AWS using various querying tools.
- Investigation of machine learning at scale using Amazon SageMaker on AWS.
- Created a POC involved in loading data from LINUX file system to AWS S3 and HDFS.

TECHNICAL SUMMARY

MIYA MA | HADOOP | 628-222-5612| weima3389@gmail.com

MIYA MA

SENIOR BIG DATA ENGINEER

SCRIPTS

SQL, Spark, Spark Sql, Spark Streaming, Spark Structured Streaming, Pig, Hive, XML, Python, Bash, Scala, R

DISTRIBUTIONS: Hortonworks, Cloudera, AWS EMR

HADOOP ECOSYSTEM

Maven, SBT, YARN, Flume, Kafka, Maven, Oozie, Pig, Spark, Tez, Zookeeper, HDFS, Kibana, Flume, Zookeeper, Tableau, Airflow

DATABASE: SQL, MySQL, Oracle, Cassandra, Hbase, Hive, MongoDB, NoSQL

CONTINUOUS INTEGRATION SERVERS

Jenkins, GitHub, Bitbucket, GitLab

SOURCE CONTROL MANAGEMENT

Git, SVN

PROFESSIONAL EXPERIENCE

SENIOR DATA ENGINEER

YELP | Washington, DC Jan 2019 – Present

- Developed Cloud-based Big Data Architecture using Hadoop and AWS which created the foundation of this Enterprise Analytics initiative in a Hadoop-based Data Lake.
- Created variation of the lambda architecture consisting of near real-time using Spark SQL.
- Apache Open source version with Mesos job scheduler.
- Developed, designed tested Spark SQL clients with Scala, PySpark.
- Created multi-node Hadoop and Spark clusters in AWS instances to generate terabytes of data and stored it in AWS HDFS.
- Deployed the application jar files into AWS instances.
- Captured and transformed real-time data from Amazon Aurora into a suitable format for Scalable analytics.
- Manipulated and analyzed complex, high volume, and high dimensional data in AWS using various querying tools.
- Connected database and Tableau, using sql query join tables and create dashboard.
- Skilled in creating and executing successful Google Analytics campaigns.
- Analyze the performance of digital strategies to yield business recommendations.

MIYA MA | HADOOP | 628-222-5612| weima3389@gmail.com

MIYA MA

SENIOR BIG DATA ENGINEER

- Able to research and propose online search and social media strategies to increase Web site traffic and conversions.
- Investigation of machine learning at scale using Amazon SageMaker on AWS.
- Reviewing Business requirement document for completeness, analyzes actual, and forecast budgeting and forecasting requirements.
- Install Hadoop using Terminal and set the configurations.
- Setup cloud compute engine managed and unmanaged mode and SSH key management.
- Hadoop data ingestion and Hadoop cluster handing in real time processing using Kafka and spark.
- Installed spark and pyspark library in terminal using CLI in bootstrapping step.
- Worked on importing the unstructured data into the HDFS using spark streaming and Kafka.
- Developed PySpark application as ETL tool.
- Processed data with natural language toolkit to count important words and generated word clouds.
- Constructed, evaluated and compared models such as regression, decision tree, random forest, bagging, boosting, PCA and clustering.
- Set business objectives and market positioning, designed a relational database in Lucid Chart.
- Involved in implementation of analytics solutions through Agile/Scrum processes for development and quality assurance.
- Conducted exploratory data analysis and managed dashboard for weekly report, using.
- Tableau Desktop connecting to Hadoop Hive tables.
- Created Hive and Impala queries to spot emerging trends by comparing data with historical metrics.
- Imported data from web service into HDFS and transformed data using Pig.

SENIOR BIG DATA ENGINEER

American Express | New York, NY Aug 2017 – Jan 2019

- Lead a team of 6 to develop an automated Hadoop installation shell script including install and start ssh, git, netcat, python, Scala, sbt, Hadoop, flume, Kafka and spark.
- Transformations using MapReduce, Hive to load data into HDFS.
- Managed and reviewed Hadoop log files.
- Performed maintenance, monitoring, deployments, and upgrades across infrastructure that supports all Hadoop clusters.
- Skilled on importing and exporting data using flume and Kafka.
- Optimized data storage in Kafka Brokers within the Kafka cluster by partitioning Kafka Topics.
- Experience in using Kafka as a messaging system to implement real-time Streaming solutions using spark steaming.
- Extracted Real time feed using Kafka and Spark Streaming and convert it to RDD.
- Processed data in the form of Data Frame and save the data as Parquet format in HDFS using spark.
- Integrated Kafka with Spark streaming for high speed data processing.

MIYA MA | HADOOP | 628-222-5612| weima3389@gmail.com

MIYA MA

SENIOR BIG DATA ENGINEER

- Imported data from different source like HDFS and api into spark RDD for further processing.
- Collected, aggregate, and move data from servers to HDFS using spark and spark streaming.
- Wrote Scala and python script using spark to read and count word frequency from 81 doc files and the generate tables compare frequency across the files.
- Worked with various stakeholders for gathering requirements to create as-is and as-was dashboards.
- Writing SQL queries for data validation of the reports and dashboards as necessary.

DATA ENGINEER

Walmart | Bentonville, AR May 2016 – Aug 2017

- Installed Kafka and start zookeeper, servers, producer and consumer in terminal.
- Built continuous spark streaming ETL pipeline with spark, Kafka, Scala, HDFS MongoDB.
- Wrote python scripts as a producer and consumer to get streaming data from twitter api and save it to HDFS.
- Used Kafka to transform live steaming with the batch processing to generate report.
- Performed streaming data ingestion process through pyspark.
- Configured spark streaming to receive real-time data from Kafka and store to HDFS using python and Scala.
- Developed custom aggregate functions using Spark SQL and performed interactive querying.
- Connected various data centers and transferred data between them using Spark and various ETL tools.
- Used Spark-SQL and Hive Query Language (HQL) for getting customer insights, to be used for critical decision making by business users.
- Learned and adapt to perform for the CICD tool (GITHUB, Jenkins) chain that is available at Customer environment or proposed to be made available.
- Accessed Hadoop file system (HDFS) using Spark and managed data in Hadoop data lakes with Spark.
- Developed new flume agents to extract data from Kafka and other web servers into HDFS.
- Requirement gathering for data warehouse.

DATA ENGINEER

Wish | San Francisco, CA Mar 2015 – May 2016

- Wrote Hive Queries for analyzing data in Hive warehouse using Hive Query Language.
- Generated informative visualizations in Tableau and delivered results through presentation to a group of 46.

MIYA MA | HADOOP | 628-222-5612| weima3389@gmail.com

MIYA MA

SENIOR BIG DATA ENGINEER

- Prepared scripts to automate ingestion of data in Python as needed through various sources such as API and save it to HDFS.
- Created modules for Spark streaming in data into Data Lake using Spark.
- Installed flume in terminal and configured the source, channel and sink.
- Configured flume for efficiently collecting, aggregating and moving large amount of data.
- Wrote a flume configuration to read from file and write the file to HDFS.
- Experience on collecting real-time data from diverse sources like webservers and social media using python and Scala and storing in HDFS for further analysis – ingested through flume.
- Imported real-time logs to Hadoop Distributed File System (HDFS) using Flume.
- Skilled in phases of data processing (collecting, aggregating, moving from various sources) using Apache Flume and Kafka.
- Experience in transferring Streaming data from different data sources into HDFS and HBase using Apache Flume.
- Worked on Spark SQL to check the data.
- Wrote Spark applications for data validation, cleansing, transformation, and custom aggregation.
- Created and delivered presentations to various business teams regarding the current status and key decisions regarding their business issues.

EDUCATION

Masters' of Science Degree in Business Analytics
University of Maryland