**Mayang Vasava**

[**Vasavamayang@gmail.com**](mailto:Vasavamayang@gmail.com)

**LinkedIn: https://www.linkedin.com/in/mayang-vasava-1415a81b0/**

**848-260-8497**

**PROFESSIONAL SUMMARY:**

* 8+ years of IT experience which includes 5+ years of work experience in Big Data, Hadoop ecosystem related technologies.
* Overall, 5 years of experience in application development and design using **Object Oriented Programming,** **Scala Spark**, **PL/SQL**, **JAVA technologies.**
* Good knowledge of Hadoop Architecture and various components such as **HDFS, Job Tracker, Task Tracker, Namenode and Datanode.**
* Experience in Hadoop ecosystem components like Map Reduce, HDFS, Hive, Scala, Sqoop, Pig, Kafka, **Nifi** for scalability, distributed computing and high-performance computing.
* Experience in the developing NoSQL database**.**
* Experience in good understanding of **Apache Storm-Kafka pipelines.**
* Extensive experience working in **SQL Server** and **MySQL** database.
* Good Experience in data loading from **Oracle** and **MYSQL** databases to **HDFS** system using **Sqoop** (Structure Data) and **Flume** (Log Files & XML).
* Knowledge on analyzing data interactively using **Apache Spark**.
* Knowledge in developing **PIG Latin Scripts** and using **Hive Query Language** for data analytics.
* Experienced in writing custom **Hive UDF's** to incorporate business logic with **Hive queries**.
* Good experience in optimizing **Map Reduce algorithms** using **Mappers**, **Reducers**, **combiners** and **partitioners** to deliver the best results for the large datasets.
* Experience in understanding the security requirements for Hadoop and integrate with **Key Distribution Centre**.
* Proficient in **Java, Scala** and **Python.**
* Expertise in Amazon AWS concepts like **EMR** and **EC2** web services which provides fast and efficient processing of Big Data.
* Hands on experience in using BI tool like **Tableau.**
* Detailed understanding of **Software Development Life Cycle (SDLC)** and sound knowledge of project implementation methodologies including **Waterfall and Agile.**
* Involved in design and development of various web and enterprise applications using various technologies like **GCP**, **XML**, **Java Script**, **HTML**, **AJAX** and **Amazon Web Services**.
* Experience in constructing pipelines using workflow tools like **Oozie**.
* Experienced in providing real time analytics on big data platforms using **HBase, Cassandra** and **Mongo DB.**
* Hands on experience in application development using core **JAVA, RDBMS, Linux shell scripting** and also developed **UNIX shell** scripts to automate various processes.
* Having Experience on Development applications like Eclipse, Intellij etc.
* Expertise in Unit Testing, Integration Testing, System Testing and experience in preparing the Test Cases, Test Scenarios and Test plans.
* Ability to work independently as well as in a team and able to effectively communicate with customers, peers and management at all levels in and outside the organization.

**AREA OF EXPERTISE:**

|  |  |
| --- | --- |
| **Programming/Scripting Languages** | Scala, PySpark, Python, SQL |
| **Big Data** | Hadoop, HDFS, Hive, sqoop, Spark, Kafka |
| **Other tools** | Microsoft Office tools, VSTS, Git |
| **Databases:** | Oracle, MYSQL, Hbase |
| **Big data Eco System** | HDFS, Oozie, Zookeeper, Spark SQL, Spark streaming, Hue, Ambari, Impala. |
| **File Formats** | Txt, XML, JSON, Parquet, ORC, CSV |
| **Cloud Computing** | AWS, Azure |
| **Visualization and Reporting Tools** | Microsoft Power BI |

**Professional Experience:**

**American Express, Phoenix, AZ Jan’22- Present**

**Role: Data Developer**

**Project:MIRA**

**Description:**

The Merchant Insight Reporting and Analytics (MIRA) plat form is a centralized and governed global merchant analytics & reporting platform. It comprises of three main components, Global Merchant Data Warehouse (GMW) and Merchant Analytics and Reporting Capability (MARC), & PDS (Tableau). The MIRA platform sources merchant data from a variety of regional transactional systems, and then synthesizes, cleanses, and governs the data to create a globally consistent merchant information repository that is easy to access and understand. MIRA integrates various Merchant aspects to provide best in class data analytics at various levels to identify trending patterns and growth opportunities for American Express and embrace compliance /regulatory needs.

**Responsibilities:**

* Design and developed and implemented data transformation to meet client’s expectation and gathering requirements from business team to convert them into technical requirement.
* Integrated **Maven** builds and designed workflows to automate the build and deploy process.
* Used Spark for interactive queries, processing of streaming data and integration with popular **NoSql** database for huge volume of data.
* Load and transform large sets of structured and semi structured data. Involved in creating **Hive** tables, loading and analyzing data using hive queries.
* Worked on Yellowbrick data warehouse like Teradata and support big-data needs which support **SQL** and scalability from terabytes to petabytes.
* Used spark and **spark-sql** to read the parquet data and create the table in hive using the **scala-API**.
* Handling ITSM INC Tickets on daily basis
* Developing **HQL** queries to retrieve the data and ingest back it into Cornerstone after applying necessary business logics.
* Performing detail Unit Testing: Creation of Unit test Plans/ scenarios, Test data preparation and log the test results.
* Provide the input to Quality Analyst for preparing the test plan for the system and UAT test phase.
* Deploying the artefacts into production using **CI/CD** process and deployment related activities and developing **UNIX** shell scripts for deployment
* Providing necessary technical support to team by giving the training Spark and Event engine.
* Involved in developing Hive DDLs to create, alter and drop tables.

**Environment:** PySpark, Spark SQL, Hive, HQL, Event Engine, Shell scripting, Jenkins, Github, HDFS, Linux.

**Walmart, Sunnyvale, CA Jan’21- Dec’22**

**Role: Hadoop Developer**

**Project: CBB Datalake.**

**Description:**

The golden record here in an ideal state will span across all entities which customers have access to - store+online+pharmacy+socialmedia+buyinghistory+WalmartPlus+marketingchannels+household+promotions/events+... The data lake will allow us to ingest and link customer (CBBID) with data from virtually any source within Walmart will taking care of other costs which come along with it. The notion that individuals/teams need to understand their respective events and how they co-relate to a customer will be completely managed by CBB which in turn will rely on the DataLake for providing a storage and access mechanism for teams to understand and develop different ways in increasing loyalty and share of wallet.

**Responsibilities:**

* Understanding and gathering requirements for W+ Members events from business team to convert them into technical requirement.
* Used apache Nifi to generate graphical representation of data transferring and flow.
* Designed, developed and implemented data transformation to meet client’s expectation
* Write a Python program to maintain raw file archival in **GCS** bucket.
* Configured, Monitored and Automated **GCP** Services as well as involved in deploying the content cloud platform using **Google Storage Buckets** and created Google storage buckets and maintained and utilized the policy management of these buckets and backup on Google cloud.
* Develop and deploy the outcome using spark and Scala code in Hadoop cluster running on GCP.
* Created environment and worked on the **Airflow**for scheduling the jobs by using the Dags and accomplished the POC which is Triggering the **Dags**.
* Used **Data prep** for converting the Raw data to the Redefined data and used Cloud Storage bucket for storing that data and from Cloud Storage exported the data to the**SQL** and the Big Query.
* Worked on Maven project to single grade project with designed workflows to automate the build process. Developed Java class to Scala class.
* Optimizing Java code, Hive scripts for better scalability, reliability and performance.
* Involved in creating Hive Tables, loading with data and writing hive queries.
* Analyzed large amounts of data sets to determine optimal way to aggregate and report on it.
* Worked with Play framework and Akka parallel processing.

**American Express, Phoenix, AZ Aug’19- Dec’20**

**Role: Hadoop Developer**

**Project: California Consumer Privacy Act.**

**Description:**

The California Consumer Privacy Act (CCPA) of 2018 was passed on June 28th 2018. The law, which will take effect on January 1st 2021, introduces new privacy rights for consumers and will force companies that conduct business in the State of California to implement structural changes to their privacy programs. The new rights given to California consumers are similar to the rights provided in the European Union’s General Data Protection Regulation (GDPR). The CCPA also subjects non-compliant businesses to expensive fines, class-action lawsuits, and injunctions.

**Responsibilities:**

* Understanding the existing process and gathering requirements from business team to convert them into technical requirement
* Developed data pipelines using ETL tool Magellan.
* Designed, developed and implemented data transformation to meet client’s expectation
* Developing HQL queries to retrieve the data and ingest back it into Cornerstone after applying necessary business logics.
* Developed CDC (Change Data Capture) process to handle upcoming incremental data.
* Developed PySpark applications to load JSON data into Hive table from CSV and Json files and create the final file and insert it into cornerstone (HDFS).
* Involved in designing the data model in Hive for migration the ETL process into Hadoop and wrote Scripts to load data into Hadoop Environment.
* Deploying the artefacts into production using **CI/CD** process.
* Providing necessary technical support to team by giving the training Spark and Event engine.
* Hive QL scripts to create load and query tables in hive and also performed sentiment analysis using hive Ql scripts.

**Environment:** PySpark, Spark SQL, Hive, HQL, Magellan Tool, Event Engine, Shell scripting, Jenkins, Github, HDFS, Linux.

**Client: PCG (Public Consulting Group), New York,NY Jul’17– Jul’19**

**Role: Hadoop/Big Data Developer**

**Description:**    
PCG (Public Consulting Group) helps primarily public sector health, education, and human services organizations make measurable improvements to their performance and processes.

**Responsibilities:**

* Worked with highly unstructured and semi structured data of 90 TB in size (270 TB with replication factor of 3)
* Developing scripts and Batch Job to schedule various Hadoop programs.
* Used **Pig** as ETL tool to do transformations, event joins, filter & some pre-aggregations.
* Collecting and aggregating large amounts of log data using **Apache Flume** and staging data in HDFS for further analysis.
* Designed the data models to be used in data intensive AWS Lambda applications which are aimed to do complex analysis creating analytical reports for end-to-end traceability, lineage, and definition of Key Business elements from Aurora.
* Used **Spark stream** processing to get data into in-memory, implemented **RDD** transformations, actions to process as units
* Created/modified **UDF** and **UDAFs** for Hive.
* Populated **HDFS** and **Cassandra** with huge amounts of data using **Apache Kafka.**
* Used DML statements to perform different operations on **Hive Tables**.
* Developed Hive queries for creating foundation tables from stage data.
* Adjusting the minimum share of maps and reducers for all the queues.
* Analyzed the data by performing Hive queries and running **Pig scripts** to study customer behavior.
* Managed **Amazon Web Services** (AWS) EC2.
* Working with **Apache Crunch** library to write, test and run HADOOP MapReduce pipeline jobs.
* Efficiently put and fetched data to/from HBase by writing Map/Reduce job in **Java/Python.**
* Cluster coordination services through **Zookeeper**.
* Creating **Hive tables**, **dynamic partitions**, **buckets for sampling**, and working on them using **Hive QL**.
* Experienced on loading and transforming of large sets of **semi structured data** using **Pig Latin operations**.
* Extracted the data from Teradata into HDFS using **Sqoop.**
* Data Visualization using **Tableau** for reporting from Hive Tables.

**Environment:** Hadoop, HDFS, Apache, Map Reduce, Hive, Flume, Sqoop, Zookeeper, Kafka, Storm, Cassandra, Spark, Linux**.**

**Client: The Travelers, Inc.,New York, NY Jun’14– Jun‘17**

**Role: Big Data Developer**

**Description:**

The Travelers Companies, Inc. is an American [insurance](https://en.wikipedia.org/wiki/Insurance) company. It is the second largest writer of U.S. commercial property casualty insurance and the third largest writer of U.S. personal insurance through independent agents. Travelers is incorporated in Minnesota, with headquarters in [New York City](https://en.wikipedia.org/wiki/New_York_City) .

**Responsibilities:**

* Have real-time experience of **Kafka-Storm** on **HDP 2.2** platform for real time analysis.
* Created **PoC** to store Server Log data in **MongoDB** to identify System Alert Metrics
* Implemented Hadoop framework to capture user navigation across the application to validate the user interface and provide analytic feedback/result to the UI team
* Loaded data into the cluster from dynamically generated files using Flume and from relational database management systems using **Sqoop**.
* Performed analysis on the unused user navigation data by loading into **HDFS** and writing **MapReduce jobs.** The analysis provided inputs to the new APM front end developers and lucent team.
* Wrote **MapReduce** jobs using **Java API and Pig Latin.**
* Loaded the data from **Teradata to HDFS** using Teradata Hadoop connectors.
* Wrote **Pig scripts** to run **ETL jobs** on the data in **HDFS** and further do **testing**.
* Used **Hive** to do analysis on the data and identify different correlations.
* Imported data using **Sqoop** to load data from **MySQL** to **HDFS** and **Hive** on regular basis.
* Written Hive queries for data analysis to meet the business requirements.
* Involved in collecting, aggregating and moving data from servers to HDFS using **Apache Flume**.
* Involved in creating **Hive tables** and working on them using **HiveQL and** perform data analysis using **Hive** and **Pig.**
* Automatically Importing data regular basis using **sqoop** to into the Hive partition by using **apache Oozie.**
* Supported Map Reduce Programs those are running on the cluster**.**
* Weekly meetings with technical collaborators and active participation in code review sessions with senior and junior developers.
* Continuous monitoring and managing the Hadoop cluster through Cloudera Manager

**Environment:** Hadoop, MapReduce, HDFS, Pig, Hive, HBase, Flume, ZooKeeper, Cloudera Manager,Oozie, Java (jdk1.6), MySQL, SQL, Windows NT, Linux