Mounika M

Portland, Oregon | (913) 353-4626 |mounika4171@gmail.com

# Professional Summary

# 6+ years of IT experience in a variety of industries, which includes hands on experience in Hadoop, Hive, Spark, SQOOP

* Expertise on coding in different technologies i.e. Python, Java Unix shell scripting.
* Extensive experience in Big Data ecosystem and its various components such as SPARK, SCALA, MapReduce, HDFS, HIVE, PIG, Sqoop, Zookeeper, Oozie and Flume.
* Well versed experience in Amazon Web Services (AWS) Cloud services like EC2, S3, EMR, DynamoDB.
* Good experience in working with concepts of Hadoop Architecture and its components such as HDFS, Job Tracker, Task Tracker, Name Node, Data Node, Secondary Name node, and MapReduce concepts.
* Experience in handling different file formats like Text files, Sequence files, Avro data files using different SerDe in Hive.
* Mastered in using different columnar file formats like RCFile, ORC and Parquet formats.
* Experience in working with MapReduce Framework and Spark execution model.
* Hands-on experience in programming with Resilient Distributed Datasets (RDDs), data frames and dataset API.
* Experienced in Partitioning, bucketing concepts in Hive and designed both Managed and External tables in Hive to optimize performance.
* Experienced in writing custom Hive UDF's to incorporate business logic with Hive queries.
* Experience in process improvement, Normalization/de-Normalization, data extraction, data cleansing, data manipulation on HIVE.
* Experience in loading data files from HDFS to Hive for reporting.
* Experience in writing Sqoop command to import data from Relational database to Hdfs.
* Having experience in SQL Server and Oracle Database and in writing queries.
* Exploring with the Spark for improving the performance and optimization of the existing algorithms in Hadoop using Spark Context, Spark-SQL, Data Frame, Pair RDD's, Spark YARN.
* Developed Pig Latin scripts to extract the data from the web server output files to load into HDFS.
* Extensively used Pig for data cleansing.
* Developed the Pig UDF's to pre-process the data for analysis.
* Developed workflow in Oozie to automate the tasks of loading the data into HDFS and pre-processing with Pig.
* Used Python scripts to build a workflow in Autosys to automate the tasks in three zones in cluster.
* Experienced working with Business team for gathering the requirements and fully understand the business requirements.
* Designed and created data extracts, supporting Power BI, Tableau, or other visualization tools reporting applications.

# Technical Skills

|  |  |
| --- | --- |
| **Hadoop Ecosystem** | HDFS, MapReduce, Apache Spark, Hive, YARN, Oozie, Zookeeper, Sqoop, HBase, AWS EMR, Cloudera |
| **Cloud Services** | Amazon web services (AWS) |
| **Versioning Tools** | GIT |
| **CI/CD tools** | Jenkins, JIRA |
| **RDBMS** | Oracle, MySQL, SQL server, Teradata |
| **No SQL** | MongoDB, Cassandra |
| **Programing Languages** | C, Python, SQL, Java, PL/SQL, Linux Shell Script |
| **Operating system** | Linux, Windows |
| **Other tools** |  MS Visual Studio, Eclipse, Putty, Win SCP |
| **Data Visualization Tool** | Tableau, Excel Pivot |

# Professional Experience

## Spark Developer | Staples, Framingham, MA | August 2019 – present

* Used Sqoop to Import data from Oracle,Teradata and several flat files to HDFS and used Apache Spark to Perform transformations and actions.
* Experience in loading the data from Hive to AWS S3 and Snowflake using spark API
* Developed ETL frameworks for data using PySpark
* Performed fine-tuning of spark applications/jobs to improve the efficiency and overall processing time for the pipelines
* Used broadcast variables in Spark, effective & efficient Joins, transformations, and other capabilities for data processing.
* Used Spark-SQL to perform event enrichment and to prepare various levels of user behavioral summaries.
* Worked with EMR, S3, and Ec2 services in AWS cloud.
* Implemented the workflows using Oozie scheduler and Control-M to automate tasks.
* Experience in forking the jobs using Oozie to execute the jobs in parallel.
* Experience in integrating the code changes to Bitbucket repository and build using JIRA CICD pipeline
* Experienced working with Spark Core and Spark SQL using Python.
* Interacted with the infrastructure, network, database, application, and BA teams to ensure data quality and availability.
* Strong experience and knowledge in Data Visualization with Tableau
* Experience in designing, developing, scheduling reports/dashboards using Tableau.

**Technologies:** HDFS, Hadoop 2.x, Hive, Sqoop, Flume, Spark 2.3.x, Storm, Kafka, MapReduce, Oozie scheduler, Oracle 11g, YARN, AWS, python, Shell Scripting, Tableau.

## DATA ENGINEER | INFOSYS | MARCH 2014 – DECEMBER 2018

**Client: BCBS Jan 2016 - Dec 2018**

**Role: Data Engineer**

**Responsibilities:**

* Responsible in developing this project from scratch using Python, Spark using python (Pyspark) in Agile model.
* Experience in hive partitioning, bucketing and perform joins on hive tables.
* Involved in designing Hive schemas, using performance tuning techniques like partitioning, bucketing.
* Used Sqoop to import the data to Hadoop Distributed File System (HDFS) from RDBMS.
* Used ETL to develop jobs for extracting, cleaning, transforming, and loading data into data warehouse.
* Involved in building the ETL architecture and Source to Target mapping to load data into Data warehouse.
* Involved in writing the shell scripts for exporting log files to Hadoop cluster through automated process.
* Used oozie job scheduler end to end data processing pipelines and scheduling the workflows.
* Developed complex logics and generic code for different validations using Spark.
* Worked on Spark SQL for table level validations and fetching meta data from static tables.
* Created job execution engine using Hive, which tracks job status for each process level, file level and at validations level.
* Trained and lead team of associates and colleagues in understanding the functionality of framework and uplifted them technically in Hadoop, Spark and Bigdata technologies.
* Used bit bucket as common repositories for code sharing.
* Maintained timely delivery in every sprint.

**Technologies:** Spark SQL, Python, HDFS, Hadoop 2.x, Hive, Sqoop, Spark 2.3.0, Cloudera, MapReduce, Oozie scheduler, Oracle 12g, YARN, python, Shell Scripting.

**Client: Westpac Mar 2014-Dec 2015**

**Role: ETL Data Analyst**

**Responsibilities:**

* Create/Enhance the JCL Jobs, Procs, Batch & Online Programs.
* Preparation of Productionized Jobs.
* Responsible for preparing Knowledge Management Docs for Team and shared across other projects which helped in reducing the efforts.
* Responsible for helping team members by giving solutions who are into same as well as other applications of our project.
* Always involved in understanding of each applications across the project by taking the cross functional knowledge.
* Experience in organizing the database and tables in Oracle database as per the requirement of the incoming source file at the staging area.
* Involved with the data ingestion team for creating different data pipeline solution for loading data from different sources (RDBMS, CSV and XML).
* Created control M job with incremental load to store the data to snapshot area in Oracle DB.
* Experience in data analysis in OLAP CUBES and in ORACLE databases
* Experience in writing SQL scripts using GROUP, JOIN, etc. operations to transform raw data from several data sources into forming baseline data.
* Actively participating in the code reviews, meetings and troubleshoot technical issues and solving them.
* Wrote calculated columns, Measures query in power bi desktop to show good data analysis techniques.

**Technologies:** ORACLE, Qlik View, TFS, Control-M, JIRA, Power BI.