Aayush Limbodia
Big Data Engineer

Aayush.limbodia95@gmail.com
214-810-5081

SUMMARY

Data Engineer with 7 years of experience as Data Engineer, Python Developer. Proficient in designing, documenting, development, and implementation of data models for enterprise-level applications. Background in Data Lake, Data Warehousing, ETL Data pipeline & Data Visualization. Proficient in Big data storage, processing, analysis, and reporting on all major Cloud vendors- AWS, Azure.

- Experience in Big Data ecosystems using **Hadoop**, **MapReduce**, **YARN**, **HDFS**, **HBase**, **HIVE**, **Sqoop**, Storm, **Spark**, **Python**, **Airflow**, **Kinesis** and **HBase**.
- Experience developing **Spark** applications using **Spark Core**, **Streaming**, **SQL**, **Data Frames**, **Datasets** & **Spark-ML**. Developed **Spark Streaming** jobs by developing RDDs using **Scala**, **PySpark** and **Spark-Shell**.
- In-depth understanding/knowledge of Hadoop Architecture and components including **HDFS**, **Job Tracker**, **Task Tracker**, **Name Node**, **Data Node** and **MapReduce**. Worked in **Design**, **Implementation**, **Deployment** and **Maintenance** of end-to-end **Hadoop** based analytical solutions.
- Experienced in **HIVE Queries** to process large sets of **structured**, **semi-structured** & **unstructured** data. Experience in loading data into **HDFS** using **Sqoop** as well as saving data in **Hive** tables.
- Involved in end-to-end implementation of **Enterprise Data Warehouse, Data Lakes** & **Data Mart** with **Batch** and **Real-time processing** using **Spark** streaming, **Kafka**, **Flume** and **Sqoop**.
- Experience in setting up workflow using **Apache Airflow** and **Oozie** to manage & schedule **Hadoop** jobs.
- Experience in ETL using **Informatica DEI**/**BDM**, **Power Center**, **DataStage** & **IICS** tools.
- Experience in **configuration, deployment** & **automation** of major Cloud environments (**AWS, Azure** & **GCP**).
- Worked with **AWS EC2** cloud instance. Used **EMR**, **Redshift**, and **Glue** for data processing.
- Worked with **AWS storage**, **OLTP**, **OLAP**, **NoSQL** & their **data warehouse AWS RedShift**.
- Worked on creating IAM policies for delegated administration within **AWS**. Configured **Users**/**Roles**/**Policies.**
- Proficient in AWS **Code Pipeline** and worked with **Code Commit**, **Code Build** & **Code Deploy**.
- Hands on experience with **Microsoft Azure Cloud** services, **Storage** Accounts and **Virtual Networks.** Worked on **Security** in **Web Applications** using **Azure** and deployed **Web Applications** to **Azure**.
- Experience in **GCP** platform**- compute engine**, **cloud load balancing**, **cloud storage**, **database** (**Cloud SQL, Bigtable, Cloud Datastore)**, **stack driver monitoring** and **cloud deployment manager**.
- Developed and designed system to collect data from multiple portals using **Kafka** and process it using **Spark**. Designed and implemented **Kafka** by configuring **Topics** in **new Kafka** cluster in all **environments**.
- Experience in **data preprocessing** (data **cleaning**, data **integration**, data **reduction**, and data **transformation**) using **Python** libraries including **Boto3** and **Pandas** for data analysis and numerical computations.
- Experience working on various file formats including **delimited text** files, **clickstream log** files, **Apache log** files, **Parquet** files, **Avro** files, **JSON** files, **XML** files and others.
- Good understanding of **data modeling** (**Dimensional** & **Relational**) concepts like **Star-Schema Modeling**, **Schema Modeling, Fact** and **Dimension tables**.
- Involved in various projects related to **Data Modeling**, **System/Data Analysis**, **Design** and **Development** for both **OLTP** and **Data warehousing** environments.
- Experienced in **Snowflake** data warehousing to provide stable **infrastructure**, **architecture**, **secured environment**, **reusable generic frameworks**, **technology expertise**, **best practices** and **automated SCBD.**
- Worked with **Snowflake** features like **clustering**, **time travel**, **cloning**, **logical data warehouse**, **caching** etc.
- Worked in **TERADATA** Database **design**, **implementation** & **maintenance** in large scale Data Warehouse. Proficient in **TERADATA SQL**, Stored Procedures, Macros, Views, Indexes Primary, PPI & Join indexes.

TECHNICAL SKILLS

ETL Tools	AWS Glue, Azure Data Factory, Airflow, Spark, Sqoop, Flume, Apache Kafka, Spark Streaming
NoSQL Databases	MongoDB, Cassandra, Amazon DynamoDB, HBase, GCP DataStore
Data Warehouse	AWS RedShift, Google Cloud Storage, SnowFlake, Teradata, Azure Synapse
SQL Databases	Oracle DB, Microsoft SQL Server, IBM DB2, PostgreSQL**,** Teradata, Azure SQL Database, Amazon RDS, GCP Cloud SQL, GCP Cloud Spanner
Hadoop Distribution	Cloudera, Hortonworks, MapR, AWS EMR, Azure HDInsight, GCP DataProc
Hadoop Tools	HDFS, Hbase, Hive, YARN, MapReduce, Pig, HIVE, Apache Storm, Sqoop, Oozie, Zookeeper, Spark, SOLR, Atlas
Programming & Scripting	Spark Scala, Python, Java, MySQL, PostGreSQL, Shell Scripting, Pig Latin, HiveQL
AWS	EC2, S3, Glacier, Redshift, RDS, EMR, Lambda, Glue, CloudWatch, Rekognition, Kinesis, CloudFront, Route53, DynamoDB, CodePipeline, EKS, Athena, QuickSight
Web Development	HTML, XML, JSON, CSS, JQUERY, JavaScript
Monitoring Tools	Splunk, Chef, Nagios, ELK
Source Code Management	JFrog Artifactory, Nexus, GitHub, CodeCommit
Containerization	Docker & Docker Hub, Kubernetes, OpenShift
Build & Development Tools	Jenkins, Maven, Gradle, Bamboo
Methodologies	Agile/Scrum, Waterfall

PROFESSIONAL EXPERIENCE

TDS Telecom, AWS Data Engineer, Wisconsin Mar 2021 – Current

Responsibilities-

- Designed, and build scalable distributed data solutions using with **AWS** & planned migration plan for existing on-premises **Cloudera Hadoop distribution** to **AWS** based on business requirement.
- Implemented simple to complex transformation on Streaming Data and Datasets. Worked on analyzing Hadoop cluster and different big data analytic tools including **Hive**, **Spark**, **Python**, Sqoop, Oozie.
- Worked **with legacy on-premises VMs** based on **UNIX** distributions. Worked with **batch data** as well as **3rd Party data** through **FTP**.
- Worked with **HDFS** that stores **distributed data**. Configured **Oozie** along with **Sqoop** to **ingest relational data**.
- Wrote YML files for **Kafka** Producers for ingesting streaming data. Assigned partitions to customers. Installed Kafka manager for consumer logs and for monitoring Kafka Metrics.
- Developed Scala scripts using both Data frames/SQL/Data sets and RDD/MapReduce in **Spark** for Data Aggregation, queries and writing data back into OLTP system through Sqoop.
- Developed **PySpark** Streaming by consuming static and streaming data from different sources.
- Used Spark Streaming to stream data from external sources using Kafka service and responsible for migrating the code base from Cloudera Platform to Amazon **EMR** and evaluated Amazon eco systems components like RedShift, Dynamo DB. Having good knowledge in **NOSQL** databases like **Dynamo DB**, Mongo DB, Cassandra. Setting up and administering DNS system in **AWS** cloud using Route53.
- Performed configuration, deployment, and support of cloud services in Amazon Web Services (AWS).
- Experienced working on cloud AWS using **EMR** Performed operations on **AWS** using EC2 instances, **S3** storage, performed RDS, analytical Redshift operations and wrote various data normalization jobs for new data ingested into Redshift by building multi-terabyte of data frame.

- Developed Oozie workflows for scheduling and orchestrating the ETL process. Involved in writing Python scripts to automate the process of extracting weblogs using **Airflow** DAGs.
- Worked on creating data pipelines with Airflow to schedule **PySpark** jobs for performing incremental loads and used Flume for weblog server data. Created **Airflow** Scheduling scripts in Python.
- Involved in migrating **Oozie** workflows to Airflow to automate data pipelines to extract data and weblogs from **DynamoDB** and Oracle databases.
- Built and configured a virtual data center in the Amazon Web Services cloud to support Enterprise Data Warehouse hosting including Virtual Private Cloud, Security Groups, Elastic Load Balancer.
- Expert in implementing advanced procedures like text analytics and processing using the in-memory computing capabilities like Apache **Spark** written in Scala. Expertized in implementing **Spark** using Python and Spark SQL for faster testing and processing of data responsible to manage data from different sources.
- Implemented data ingestion and handling clusters in real time processing using Kafka.
- Developed Spark Programs using Java API's and performed transformations and actions on Data Frames.
- Spark applications using **Spark-SQL** in EMR for data extraction, transformation, and aggregation from multiple file formats for analyzing & transforming the data to uncover insights into the customer usage patterns.
- Working experience with data streaming process with Kafka, Apache **Spark**, **Hive**.
- Analyzed the **SQL** scripts and designed the solution to implement using Python.
- Designed both 3NF data models for OLTP systems and dimensional data models using star and **snowflake** Schemas.
- Developed parallel reports using **SQL** and **Python** to validate the daily, monthly, and quarterly reports.
- Designed and Developed Real time Stream processing Application using Spark, Kafka, Scala, and **Hive** to perform Streaming ETL.
- Created **PySpark** code that uses Spark SQL to generate data frames from Avro formatted raw layer and writes them to data service layer internal tables as orc format.
- Developed **Airflow** DAGs in python by importing the Airflow libraries.
- Utilized **Airflow** to schedule automatically trigger and execute data ingestion pipeline.

Environment: Hadoop, HDFS, Hive, MapReduce, Pig, PySpark, Kafka, Python 3.6, AWS (Glue, Lambda, Step Functions, SQS, Code Build, Code Pipeline, Event Bridge, Athena), Unix/Linux Shell Scripting, PyCharm, SQL Server, Oracle, Oozie

CTDI, Big Data Engineer, Texas **Jul 2020 – Dec 2020**

I managed the data architecture of a critical project. I worked with the on-premises architecture and adjacent tools, while also handling web and application servers deployed on AWS. I also worked with their Sales Team, where I designed, build & maintained their AWS Architecture.

Responsibilities-

- Worked to **design tables** in **Hive** using **SQOOP**. Implemented data processing on large datasets of different forms including structured, semi-structured and unstructured data and loaded the data in **HDFS**.
- Created multiple data processing tasks using **PySpark** that included reading data from external sources, merge data, perform data enrichment and load in to target data destinations.
- Worked on **data pipeline** to process **large** set of **data** and **configured Lookup's** for **Data Validation** and **Integrity**. Developed multiple **MapReduce jobs** in **Python** and **Pyspark data frames** for **data cleaning** and **preprocessing**.
- Wrote **Kafka producers** to stream data from external **REST APIs** to **Kafka topics**. Wrote **Spark**-**Streaming** applications to **consume** the **data** from Kafka topics and wrote Spark Python code to process the stream.
- Processed **Kafka** streams using **Pyspark** on **Databricks** and saved processed data to **Synapse Analytics**.
- Used **Spark-SQL** to load **JSON** data and create Schema RDD and **loaded** it into **Hive Tables & Cassandra.**

- Experienced in handling **large datasets** using **partitions**, Spark in-memory capabilities, Broadcasts in **Spark**, effective & efficient **Joins**, **Transformation** and other during **ingestion** process itself.
- Worked on tuning **PySpark** applications to set **Batch Interval** time, level of **Parallelism** and **memory tuning**.
- Implemented **near-real time data processing** using **Stream Sets** and **Spark** framework.
- Developed **PySpark** jobs using **Python** in the **test environment** for faster **data processing** and used **Spark SQL** for **querying**.
- Design **star schema** in **Snowflake**. Created **Snowflake** authorized views for **exposing** data to other teams.
- Write a **Python** program in **Airflow** to maintain raw file **staging**/**processing** & **archival** in **GCS** bucket.
- Helped deploy **EMR** Service that can run **Apache Spark** jobs & integrated **cloud storage** to the **clusters**.
- Wrote a **program** to **download SQL** dump and load to **AWS S3** for **Data Lake** that pulled information from **servers** to perform **Hive** tasks. Built **PySpark** based **configurable** framework to connect common Data sources like **MySQL**, **Oracle**, **Postgres**, **SQL Server** and load it in **Snowflake**.
- Build a program with **Python** and execute it in **EMR** to run **Data validation** between **raw source file** and **Snowflake** target tables.
- Used **Airflow** to build a **task orchestrator** on **AWS** that can schedule jobs in **data pipeline**.
- Coordinated with team and developed framework to generate Daily **ad hoc** reports and extract data from various enterprise servers using **PySpark**.

Environment: Hadoop (HDFS, HBase, Hive, YARN, MapReduce, Pig, HIVE, Storm, Sqoop, Oozie, Zookeeper, Spark), Oracle DB2, AWS

Mckesson, Data Engineer, Texas May 2019 – Jun 2020

I was a part of the team that handled the data architecture. I worked with Python, Java Applications, Hadoop & a variety of ETL tools including Informatica & Apache Sqoop/Flume.

Responsibilities:

- Analyzed, designed, and build scalable distributed data solutions using with **Hadoop**, **AWS** & **GCP**.
- Worked on multi-tier applications using **AWS** services (**EC2**, **Route53**, **S3**, **RDS**, **Dynamo DB**, **SNS**, **SQS**, **IAM**) focusing on high-availability, fault tolerance, and auto-scaling in **AWS Cloud Formation**.
- Conducted data cleansing for unstructured dataset by applying **Informatica Data Quality** to identify potential errors and improve **data integrity** and **data quality**.
- Prepared sources (database, flat files) to connect with **Informatica PowerCenter** and created file list by using **UNIX** to improve files processing efficiency.
- Participated in documenting **Data Migration** & **Pipeline** for smooth transfer of project from development to testing environment and then moving the code to production.
- Generated server-side **PL/SQL** scripts for **data manipulation** and **validation** and **materialized views** for remote instances.
- Developed **PL/SQL triggers** and **master tables** for automatic creation of primary keys.
- Created PL/SQL **stored procedures, functions, and packages** for moving data from staging area to data mart.
- Used **Data Frame API** in Scala to convert distributed data into named **columns** & helped develop **Predictive Analytics** using **Apache Spark Scala APIs**.
- Developed **Scala** scripts using both **Data frames**/**SQL**/**Data sets** and **RDD**/**MapReduce** in **Spark** for **Data Aggregation**, **queries** and writing data back into **OLTP** system through **Sqoop**.
- Developed **Hive queries** to **pre-process** the data required for running business processes.
- Worked on **Hive** queries and **Python Spark SQL** to create **HBase** tables to load large sets of structured, semi-structured and unstructured data coming from **UNIX**, **NoSQL** databases, and a variety of portfolios.

- Loaded data into **Spark RDD** and in-memory data computation to generate the **output** response stored datasets into **HDFS**/ **Amazon S3** storage/ relational databases.
- Worked on tuning **Spark** applications to set **Batch Interval** time, level of **Parallelism** and **memory tuning**.
- Implemented **near-real time data processing** using **Stream Sets** and **Spark** framework.
- Developed **Apache Spark** jobs using **Python** in the **test environment** for faster **data processing** and used **Spark SQL** for **querying**.
- Used **Hadoop Spark Docker** container for **validating** data load for **test**/ **dev**-**environments**.
- Worked on **ETL pipeline** to **source tables** and to deliver calculated ratio data from **AWS** S3 to **Snowflake**.
- Implemented **multiple** generalized solution model using **Google AutoML**.
- Extensive expertise using the core **Spark APIs** and processing data on an **Dataproc** cluster.
- Experience in moving data between **GCP** and **AWS** using Google **Data Fusion**.
- Hands-on experience in GCP- **Compute Engine**, **Cloud SQL**, **Data Store**, **BigQuery**, **Pub/Sub**, and **DataProc**.
- Worked with building **SSH** tunnel to **Google DataProc** to access to **yarn manager** to monitor **spark jobs**.
- Developed **job processing** scripts using **Oozie** workflow. Experienced in **scheduling** & job **management**.
- Worked in development of scheduled **jobs** using with **commands**/**BASH Shell** in **UNIX**.

PrimeroEdge, Data Analyst, Texas **Sep 2018 – Apr 2019**

Involved in Python web application development (both front-end & backend), database management, testing, and deployment. Also worked with UNIX Bash Shell Scripting for job scheduling & maintenance.

Responsibilities:

- Created **APIs**, Database Model and Views Utilization using Python to build responsive web page application.
- Worked on a fully automated continuous integration system using **Git**, **Gerrit**, **Jenkins**, **MySQL** and in-house tools developed in **Python** and **Bash**.
- Participated in the SDLC of a project including **Design**, **Development**, **Deployment**, **Testing** and **Support**.
- Deployed & troubleshoot applications used as a data source for both customers and internal service team.
- Wrote and executed **MySQL** queries from Python using Python-MySQL connector and MySQL dB package.
- Implemented **UI standards** & **guidelines** in website development using **CSS, HTML, JavaScript,** and **jQuery**.
- Worked on a **Python/Django** based web application with **PostgreSQL** DB and integrated with third party email, messaging & storage services.
- Developed GUI using webapp2 for dynamically displaying the test block documentation and other features of Python code using a web browser.
- Involved in design, implementation and modifying **back-end Python code** and **MySQL database schema**.
- Developed user friendly graphical representation of item catalogue configured for specific equipment.
- Used **BeautifulSoup** for **web scrapping** to extract data & generated various capacity planning reports (graphical) using Python packages like **NumPy, matplotlib**.
- Automated different workflows, which are initiated manually with Python scripts and UNIX shell scripting.
- Fetched Twitter feeds for certain important keyword using **Twitter Python API**.
- Used **Shell Scripting** for UNIX Jobs which included Job scheduling, batch-job scheduling, process control, forking and cloning and checking status.
- Monitored **Python scripts** that are run as daemons on **UNIX** to collect trigger and feed arrival information.
- Used **JIRA** for bug & issue tracking and added algorithms to application for data and address generation.

Environment: Python 2.7 (BeautifulSoup, NumPy, matplotlib), Web Development (CSS, HTML, JavaScript, JQuery), Database (MySQL, PostgreSQL), UNIX/Linux Shell Script, JIRA, Jenkins, Git & GitHub

Data Analyst, India. Jul 2016 – June 2018

- Developing **dashboards** to track productivity and expedite remediation of issues.
- Write data definition language or data manipulation language **SQL** commands.
- Creating and executing queries utilizing various data sources to provide business information.
- Installing **SQL Server DB and power BI,** moved customer data given in CSV format into SQL Server DB.

Environment: HIVE, SQL Server, Power BI, KAFKA, Spark, Scala, R-Programming, Python-Data Structures.

Jr. Data Analyst, India May 2015 – Jul 2016

- Develop and manage data databases that support performance improvement.
- Develop and manage reports on multiple key performance indicators and metrics across Revenue Cycle Management
- Develops and evaluates network performance criteria and measurement methods.
- Assist Managers in identifying capabilities and Processes that drive continuous Improvement
- Analyze our game data by cohort to provide suggestion to the marketing team to improve the performance of our acquisition campaign

EDUCATION:

Master of Science in Finance (STEM), University of Texas at Dallas