10+ years of demonstrated experience in IT industry with expert-level skills in Big Data Hadoop Ecosystem, Apache Spark, PySpark, Scala, Python, Kafka, Data Warehousing, Data Lakes.

## Professional summary

* Solid experience in building **Data Lake** and **Data Warehouse** using **AWS** and **Google Cloud**.
* Expert in building **Batch** and **Streaming Data Pipelines** using cutting edge big data technologies and cloud services.
* Expert level knowledge on Hadoop Distributed File System (HDFS) architecture and YARN.
* Used **AWS Services** like **AWS S3, Lambda, RDS, Redshift, Glue, EMR, Ec2, Athena, IAM, StepFunction, QuickSight,** etc.
* Used **GCP Services** like **GCP Cloud storage, BigQuery, Cloud Composer, DataFlow, DataProc, Cloud Function,** etc.
* Experienced in Hive partitioning, bucketing, optimization code through set parameters and perform different types of joins on Hive tables.
* Experienced in importing and exporting data from different databases like **SQL Server, Oracle, Teradata,** etc.,
* Designed and implemented **API** endpoints for data retrieval, data ingestion, and data processing, adhering to industry best practices and security standards.
* Experienced in automating **Apache Airflow** and **Step Function.**
* Experience in Zookeeper configuration as to provide cluster coordination services.
* Extensive experience in creating RDD’s and Data Sets in **Spark** from local file system and **HDFS**.
* Developed **robust and efficient RESTful APIs using Scala**, enabling seamless integration between different data services and applications.
* Leveraged **Akka HTTP framework** in Scala to build reactive APIs that efficiently handled concurrent requests and ensured low-latency data access.
* Hands on experience in writing different **RDD** (Resilient Distributed Datasets) transformations and actions using **PySpark**.
* Created Data Frames and performed analysis using **Spark SQL** and used **RDD** and DF APIs to access variety of data sources using **Scala, PySpark, Pandas** and **Python**.
* Excellent knowledge on **Spark** core architecture.
* Acute knowledge on **Spark Streaming** and **Spark Machine Learning** Libraries.
* Created **ERM** transient and long running clusters and in **AWS** for data processing (ETL) and log analysis.
* Deployed various Hadoop applications in **EMR** - **Hadoop, Hive, Spark, HBase, Hue, Glue, Oozie** and **Presto** etc. based on the needs.
* Experience in integrating Hive with AWS S3 to read and write data from and to S3 and created partitions in Hive
* Extensively worded on ETL/Data pipelines to transform data and load from **AWS S3** to **Snowflake** and **Redshift**.
* Extensively utilized **EMRFS** (Elastic Map Reduce File System) for reading and writing SerDe data between **HDFS** and **EMRFS**.
* Experience in using Presto on **EMR** to query different types of data sources **RDBMS** and **NoSQL** databases.
* Experience in creating ad-hoc reporting, development of data visualizations using enterprise reporting tools like Tableau, Power BI, Business Objects, etc.,

## Technical skills

|  |  |
| --- | --- |
| Big Data Tools | Hadoop, Hive, Apache Spark, PySpark, HBase, Kafka, YARN, Sqoop, Impala, Oozie, Pig, Map Reduce, Zookeeper and Flume |
| Hadoop Distributions | EMR, Cloudera, Hortonworks. |
| Cloud Services | AWS – EC2, S3, EMR, RDS, Glue, Presto, Lambda, RedShift and Azure – Data Lakes, BLOB  GCP – Cloud Storage, BigQuery, Compute Engine, Cloud Composer, Data Proc, Data Flow, Pub-sub |
| BI and Data Visualizations | ETL -Informatica, SSIS, Talend, Tableau and Power BI |
| Relational Databases | Oracle, SQL Server, Teradata, MySQL, PostgreSQL and Netezza |
| No SQL Databases | Cassandra, MongoDB and HBase |
| Programming Languages | Scala, Python and R |
| Scripting | Python and Shell scripting |
| Build Tools | Apache Maven and SBT, Jenkins, Bitbucket |
| Version Control | GIT and SVN |
| Operating Systems | Unix, Linux, Mac OS, CentOS, Ubuntu and Windows |
| Tools | PUTTY, Putty-Gen, Eclipse, IntelliJ and Toad |

## Professional Experience

|  |
| --- |
| **Fonteva, Arlington, VA                                                                                          Dec 2021 - Till Date**  **Big Data Engineer** |

**Responsibilities:**

* Designed and developed batch and streaming data pipelines to load data into Data Lake and Data Warehouse. Using services like **AWS S3, RedShift, Glue, EMR, Lambda, Step Function, IAM, QuickSight, RDS, CloudWatch events** and **Trails**, etc.,
* Developed **APIs** to access enterprise metadata platform for registering datasets and migrated all the data pipeline applications from legacy to new platform.
* Developed and modified all our data application using python
* Extensively used **PySpark**, **Python** for building and modifying existing pipelines.
* Worked on creating panda’s data frames and spark data frames where needed in order optimize the performance.
* Developed **robust and efficient RESTful APIs using Scala**, enabling seamless integration between different data services and applications.
* Designed and implemented **API** endpoints for data retrieval, data ingestion, and data processing, adhering to industry best practices and security standards.
* Leveraged **Akka HTTP framework** in Scala to build reactive APIs that efficiently handled concurrent requests and ensured low-latency data access.
* Collaborated with the front-end development team to define **API contracts** and data formats, ensuring smooth communication and integration between frontend and backend systems.
* Created/modified data pipeline applications using **spark**, python, **PySpark**, **AWS EMR**, **S3**, **Lambda**, **pandas**, **SparkSQL**, **Glue**, **Pestro** and **Snowflake**
* Extensively worked on **EMR** and **Spark** **jobs** and created **IAM** roles, **S3** buckets policies based on the enterprise request.
* Created complex **SQL** queries using **SparkSQL** for all data transformation when loading data into **Snowflake**
* Experience in working **Kafka** steaming applications for **real** time data needs.
* Utilized enterprise scan tool to identify data that is sensitive in all **S3** buckets that we own and encrypt the data before downstream can use
* Deploy application to **Docker** Container using Jenkins and trigger jobs to load data from **S3** to Snowflake tables or **Snowflake** to **S3** based on the user requirements
* Collaborated with **Data Analysts** to understand the data and their end requirements and transform the data using **PySpark**, **SparkSQL** and created data frames
* Extensively worked on file formats like **Parquet**, **AVRO**, **CSV** and **JSON** files
* Experience in using **POSTMAN** to test the end points for various applications and validate the schema, data types, formats of file etc.
* Monitor applications using **PagerDuty**, and slunk logs based the data loads daily, weekly and monthly
* Experience working on Docker Images, update docker images as necessary
* Worked on Databricks, created managed clusters and performed data transformations.

Environment: PySpark, Python, Scala, Akka HTTP framework, AWS- EMR, S3, Lambda Functions, IAM, Security Groups, Glue, Presto, Git, Jenkins, Docker, Snowflake, SQL, DataFrames, pandas, Containers, PagerDuty, Splunk and Databricks

|  |
| --- |
| **USPS, Washington, D.C. July 2020 – Nov 2021**  **Data Engineer** |

**Responsibilities:**

* Designed, developed, and deployed **Data Lake** and **Data Warehouse** in **AWS Cloud**.
* Worked on creating data pipelines using **python** and **Apache airflow** integrating various data
* Processed data ingestion using **Spark** **APIs** for loading data into data lakes and **AWS S3.**
* Find data availability and spin off process data to **S3** in Parquet or Avro file formats and store on **S3** and data lake.
* Configured and development of Hadoop environment with **AWS – EC2, EMR, S3, Glue, Pestro, Redshift, Athena** and **Kinesis.**
* Created batch scripts to fetch data from **AWS S3** storage and performed Spark operations- Transformations and Actions using **Scala** and **Spark/PySpark**
* Performed **ETL** tasks using **AWS Glue** to move data from various sources to **S3** and configures crawlers to handle the **schemas**.
* Proficient in developing robust **RESTful APIs using Scala**.
* Extensive experience with **Apache Spark** for big data processing.
* Designed and implemented data warehousing solutions with **Scala and Spark.**
* Hands-on expertise in building reactive **APIs with Akka HTTP in Scala.**
* Worked on **Apache Spark** to read file formats like **Parquet**, converted **HiveQL** queries into **Spark** transformations using **Spark RDD** and **Data Frames** in **Scala**
* Extensively used **Spark** core – **Spark Context**, **Spark SQL** and **Spark Streaming** for real time data
* Performed schema level transformations using **Apache Spark** and **Scala** in **Databricks** environment for downstream users.
* Collaborated with **QA** and Data Scientists to support system, integration and user acceptance testing
* Worked on **Apache Kafka** and **Storm** for providing real-time analytics
* Extensively worked with Streaming data with **Kafka** by importing XML data for real time streaming of data into HBase and **HDFS file system**
* Utilized snappy compression techniques with Avro, Parquet files to leverage storage in HDFS
* Extensively worked on Hive to create value added procedures and wrote Hive UDF (User Defined Functions) to make the function reusable for different models.
* Implemented Partitioning and Bucketing in **Hive** tables for query optimization
* Worked on Impala for creating views for business use-case requirements on top of Hive Tables.
* Extensively worked on Window functions and Aggregate functions in Hive
* Developed Shell and Python scripts to automate data loads and data validations

**Environment:** Apache Spark, PySpark, Hive, Scala, Databricks, Azure Data lakes, Azure Blob, Snowflake, AWS S3, EMR, Glue, Pestro, Hadoop, GIT, Eclipse, Scala, Python and Apache Airflow.

|  |
| --- |
| **Vanguard, Charlotte, NC Oct 2018 – June 2020**  **Data Engineer** |

**Responsibilities:**

* Migrated on-perm **HDFS** ecosystem to **AWS Cloud**.
* Migrated the data lake and Data Warehouse from **HDFS** and **Hive** to **AWS S3** and **Redshift.**
* Analyzed large and critical data sets using Cloudera, HDFS, HBase, Hive, UDF, Sqoop, YARN, PySpark and Apache Spark
* Worked with unstructured and semi structured data of 2PB in size and developed multiple Kafka producers and Consumers to customize partitions to get optimized results.
* Worked **AWS EMR** and **S3** and processed data directly in S3 and imported into **HDFS** on EMR cluster and utilized Spark for analysis.
* Developed Hive scripts, Unix shell scripts and **PySpark** and Scala Programming for all ETL loading processes and converting the files into Parquet format in **HDFS Filesystem**.
* Developed Spark Operations which are used to Stream, Transformations, Aggregations and generating daily snapshot of customer data.
* Created **Apache Spark RDD’s** (Resilient Distributed Data sets) using Scala programming for analyzing large data sets using filters, maps, flat-maps, count, distinct, etc.
* Extensively used Sqoop to import/export data between RDBMS and hive tables, incremental imports.
* Leveraged Snappy compression format technique with Avro, Parquet file formats for storage in **HDFS**.
* Worked on **Hive tables** to created value added procedures and wrote **UDF** (User Defined Functions) to make the function reusable.
* Developed **Partitioning and Bucketing** techniques in Hive tables for query optimization for Enterprise Business Intelligence and Data Analytics teams.
* Worked on Impala query engine for data store on distributed cluster for various business user cases on top of **Hive table.**
* Streamlined **Hadoop Jobs** and workflow operations using **Oozie** and Scheduled through Autosys on monthly and quarterly basis.
* Extensively used built tools **MAVEN, SBT and IntelliJ IDEA** for building scripts.
* Built Real-time and batch pipelines using **GCP services.**
* Modeled and developed a new data warehouse from scratch and then migrated that data warehouse to **BigQuery**. Automated data pipelines and quality control checks.
* Architected and implemented a solution to migrate the data platform from **Hadoop** to **Google cloud Platform.**
* Worked on developing continuous integration (CI), continuous delivery (CD), and continuous training (CT) for the ML system using **Cloud Build.**
* Worked heavily with the business to determine reporting requirements and to explore the data with each other.
* Used Cloud functions, **DataFlow**, Cloud **pub-sub** and **BigQuery** to build streamline dashboard to monitor services.
* Created jobs using **Cloud composer** (**Airflow DAG**) to migrate data from **Data Lake** (cloud storage) to transform it using DataProc and ingest in **BigQuery** for further analysis.
* Member of the Enterprise Architecture Committee and presented multiple times to this group and other business leaders.
* Built batch and streaming jobs using **GCP** Services like **BigQuery**, **Pub/Sub**, **DataProc**, **Dataflow**, **Cloud Run**, **Compute engine** and **Cloud Composer**.
* Created Data pipelines to read data from single and multiple sources using **DataFlow** and saved data into **Cloud Storage** and **BigQuery**
* Helped to implement **PowerBI** in the organization. Developed PowerBI reports for every department in the company.
* Developed **PySpark** scripts for ETL using **DataProc**.
* Good experience with **GCP Stack driver** Trace, Profiler, Logging and Error Reporting, Monitoring
* Setup and Implement Continuous Integration and Continuous Delivery (CI & CD) Process stack using   **GIT**, **Jenkins**
* Experience in developing APIs in cloud.
* Good Knowledge on **IAM** rules and **Cloud Security.**
* Designed **PowerBI** data visualizations using cross tabs, maps, scatter plots, pie charts etc.

**Environment:** Apache Spark, AWS EMR, S3, RedShift, Oozie, Sqoop, Scala, Python, Java, MAVEN, SBT, IntelliJ, MySQL, HDFS file system, Spark RDD, PySpark, Cloudera, Horton Works, Hadoop Ecosystem, Oracle, Informatica, ETL, Pig, HBase, SQL Server, Eclipse, Pig, Hive, Shell scripting, MapReduce jobs

|  |
| --- |
| **ICICI BANK, MUM, IND                                                                    July 2015 - May 2018**  **Database Developer** |

**Responsibilities:**

* Responsible for extracting data from various flat files using SQL scripts and ETL, validate the data and load data into Datawarehouse.
* Worked on UNIX Shell scripts to extract data from various sources such as DB2 and Oracle
* Performed data analysis on both **OLATP** and **OLAP** data using complex SQL queries and delivered data solutions
* Extensively wrote **SQL queries**, **created views**, **Stored Procedures** in **Oracle database**
* Extracted data from multiple databases using database links by writing complex **pl/sql** queries to eliminate duplicate data
* Developed complex and ad-hoc reports using multiple data providers for patient services and commercial teams
* Created **ETL workflows** to extract data from **CRM** applications and loaded into Oracle Datawarehouse.
* Involved in designing physical and logical data model using **ERwin Data modeling tool.**
* Designed the relational data model for operational data store and staging areas, Designed Dimension & Fact tables for data marts.
* Extensively used **ERwin data modeler** to design Logical/Physical Data Models, relational database design.
* Created **Stored Procedures, Database Triggers, Functions and Packages** to manipulate the database and to apply the business logic according to the user's specifications.
* Created **Triggers, Views, Synonyms and Roles** to maintain integrity plan and database security.
* Creation of **database links** to connect to the other server and access the required info.
* Integrity constraints, database triggers and indexes were planned and created to maintain data integrity and to facilitate better performance.
* Used Advanced Querying for exchanging messages and communicating between different modules.
* System analysis and design for enhancements Testing Forms, Reports and User Interaction.

**Environment:** Relational Databases- MySQL, Oracle 9i, SQL\* Plus, PL/SQL, ERwin, TOAD, Stored Procedures, SQL server, Oracle, DB2, JIRA, SQL, PL/SQL, Informatica (ETL) and BI tools.

|  |
| --- |
| **Bajaj Allianz General Insurance Company Ltd , PUNE, IND MARCH 2013  - JUNE 2015**  **Database Developer** |

**Responsibilities:**

* Developed backup restore scripts for **SQL Server** as needed.
* Involved in Creation of Reports, Created **ODBC Connections** and database Connections.
* Wrote complex **ETL scripts using SQL server** integration services.
* Deeply involved in developing **SSIS packages** for extracting the data from database transforming it and loading into the data warehouse.
* Worked on various **SSIS objects** like Data flow components, Control Flow elements, Connection Manager, Event Handling and log providers.
* Based on the required reports, identified various data sources and established the connections using **SSIS** and constructed data source view.
* Formatted reports for various Output formats such as PDF, EXCEL and WORD for the convenience of the end users.
* Develop and deliver dynamic reporting solutions using **SQL Server** 2012 Reporting Services (SSRS).
* Created Drill down and Drill through reports using Crystal Reports.
* Provided **DBA assistance** in SharePoint 2007 to SharePoint2010 conversion.
* Designed and implemented data access stored procedures and triggers for automating tasks.
* Involved in job scheduling and alerts.
* Designed and implemented user login and security.
* Wrote **Parameterized Queries** for generating Tabular reports and Sub reports using Expressions, and Functions, Sorting the data, Defining Data sources and Subtotals for the reports using **SSRS 2008R2.**
* Defined report layouts including report parameters and wrote queries for drill down reports using **SSRS 2008R2**.
* Developed **complex reports** using multiple data providers, user defined objects, aggregate aware objects, charts, and synchronized queries.
* Administered interface to organize reports and data sources, schedule report execution and delivery, and track reporting history using **SSRS 2008R2**.

**Environment:** - Microsoft SQL Server 2005 Enterprise Edition, Microsoft Office SharePoint Server 2007/2010, Crystal reports, T-SQL,