**Name: Sai Kiran Gattu**

**Sr. Data Engineer**

**Email: saikiran240999@gmail.com**

**Phone: +1 848-274-4548**

**Professional Summary:**

* 10+ years of extensive experience in Data Engineering with strong expertise in Data Ingestion, Modelling, Querying, Processing, Analysis and implementation of enterprise level systems.
* Hands-on experience in the entire project life cycle including data infrastructure building, Cleansing, Manipulation, Engineering, Modelling, Optimization (performance tuning), QA and data pipeline (ETL) to Data warehouse Deployment On-premise and on Cloud (Azure, AWS).
* Well conversant with software data analytics methodologies including Data Migration, Data Transformation, Data Extraction, Data strategy, Data planning, Data scenarios, Test cases, Data Analytical Reports for manual and automation of Data Transformation.
* Designed and developed scalable data pipelines using Cloud Dataflow, Cloud Composer (Apache Airflow), and Cloud Pub/Sub for real-time and batch data processing.
* Experience in statistical programming languages like Python, SAS, Apache Spark and PySpark.
* Experience in analysing data using Python, R, SQL, Microsoft Excel, Hive, PySpark, and Spark SQL for Data Mining, Data Cleansing and Machine Learning.
* Managed GCP IAM roles, security policies, and networking components (VPC, Firewall Rules) to ensure secure and compliant data pipelines.
* Led the migration of on-premise data infrastructure to GCP, optimizing for scalability, performance, and cost using services like BigQuery, Cloud Storage, and Cloud SQL.
* Designed and maintained ETL/ELT pipelines in Palantir Foundry, integrating data from cloud platforms (AWS S3, Azure Blob), APIs, and on-prem databases.
* Experience in developing predictive models on large-scale datasets to address various business problems through leveraging advanced statistical modelling, machine learning, and deep learning.
* Good experience in synthesizing Data Engineering, Machine Learning and Big Data technologies into integrated solutions.
* Built and maintained BigQuery datasets, partitioned tables, and materialized views for optimal performance and cost efficiency.
* Automated data orchestration workflows using Cloud Composer, improving pipeline reliability and reducing manual intervention by 80%.
* Developed scalable and reusable SnapLogic pipelines (Snaps) to integrate structured and semi-structured data from REST APIs, databases, and cloud sources like AWS S3, Azure Blob, and Salesforce.
* Set up CI/CD pipelines for data workflows using Cloud Build and Terraform, ensuring version control and infrastructure as code (IaC).
* Implemented Unity Catalog across multiple workspaces in Databricks to centrally manage metadata, access controls, and lineage across teams and projects.
* Experience in development and design of various scalable systems using Hadoop technologies in various environments using Hive, Spark, HDFS, Map Reduce, YARN, Kafka, Pig, Hive, Sqoop, HBase, Oozie, Zookeeper, Cloudera Manager and Horton Works.
* Ensured multi-tenant data isolation and access controls in shared lakehouse architecture by implementing secure view patterns with Unity Catalog.
* Led migrations from legacy data platforms to Palantir Foundry, ensuring minimal downtime and data integrity.
* Experience in AWS services like VPC, Cloud front, EC2, ECS, EKS, Elastic Beanstalk, Lambda, S3, Storage gateway, RDBS, DynamoDB, Redshift, Elastic Cache, DMS, SMS, Data Pipeline, IAM, WAF, Artifacts, API gateway, SNS, SQS, SES, Auto Scaling, Cloud Formation, Cloud Watch, and Cloud Trail.
* Partnered with cross functional teams across the organization to gather requirements, architect, and develop proof of concept for the enterprise Data Lake environments like MapR, Cloudera, Hortonworks and AWS.
* Experience working with systems processing event based logs, user logs, machines, Azure AD, Azure Data Factory, event hub and event Queue etc.
* Integrated Palantir Foundry with DevOps CI/CD pipelines, ensuring automated deployment, version control, and continuous integration.
* Experience in building the Orchestration on Azure Data Factory for scheduling purposes.
* Strong experience writing, troubleshooting and optimizing Spark scripts using Python, Scala.
* Good experience with Python Libraries (Pandas, Scikit-learn, Tensorflow, Pytorch, Matplotlib etc.)
* Hands-on experience building PySpark, Spark Java and Scala applications for batch and stream processing involving Transformations, Actions, Spark SQL queries on RDD’s, Data frames and Datasets.
* Experienced in using Kafka as a distributed publisher-subscriber messaging system.
* Strong experience on performance tuning of Hive queries and troubleshooting various issues related to Joins, memory exceptions in Hive.
* Good understanding of partitioning, bucketing concepts in Hive and designed both Managed and External tables in Hive.
* Designed end-to-end data ingestion and transformation flows using SnapLogic for enterprise-level data warehousing and analytics initiatives.
* Expert in data engineering and building ETL pipelines on batch and streaming data using PySpark, Spark SQL
* Experience in building and optimizing Big Data pipelines, architectures and data sets on building Data Lakes.
* Hands-on experience in importing and exporting data between HDFS and Relational Databases using Sqoop.
* Experience in real time analytics with Spark Streaming, Kafka and implementation of batch processing using Hadoop, Map Reduce, Pig and Hive.
* Experienced in building highly scalable Big-data solutions using NoSQL column-oriented databases like Cassandra, MongoDB and HBase by integrating them with Hadoop Cluster.
* Extensive experience working on ETL processes consisting of data transformation, data sourcing, mapping, conversion and loading data from heterogeneous systems like Flat Files, Excel, Oracle, Teradata, MSSQL Server.
* Proficient at writing MapReduce jobs and UDF’s to gather, analyze, transform, and deliver the data as per business requirements and optimize the existing algorithms for best results.
* Experience working with Data warehousing concepts like Star Schema, Snowflake Schema, DataMarts, and Kimball Methodology used in Relational and Multidimensional Data Modelling.
* Experienced with version control systems like Git, GitHub, CVS, and SVN to keep the versions and configurations of the code organized.
* Created monitoring dashboards and alerts with GCP Stackdriver (Cloud Monitoring and Logging) to proactively manage data pipeline health and system performance.
* Implemented data ingestion from external APIs and cloud sources into GCP BigQuery via Cloud Functions and Cloud Dataflow, supporting near real-time analytics.
* Built end-to-end data pipelines on GCP using Cloud Dataflow, Cloud Pub/Sub, and BigQuery to process real-time and batch data with high throughput and low latency.
* Strong experience leveraging different file formats like Avro, ORC, Parquet, JSON and Flat files.
* Sound experience on Normalization and De-normalization techniques on OLAP and OLTP systems.
* Experience with Jira, Confluence and Rally for project management and Oozie, Airflow Scheduling Tools.
* Experienced in scripting skills in Python, Scala and UNIX Shell.
* Experience in building interactive dashboards, performing ad-hoc analysis, generating reports and visualizations using Tableau and PowerBI.
* Expert in importing and exporting data from different Relational Database Systems like MySQL and Oracle into HDFS and Hive using Sqoop.
* Experienced in writing SQL Queries, Stored procedures, functions, packages, tables, views, triggers using relational databases like Oracle, DB2, and MySQL, Sybase, PostgreSQL and MS SQL server.
* Good experience working in Agile/Scrum methodologies, communication with scrum calls for project analysis and development aspects.
* Experience in possessing an extensive mathematical, analytical skill with strong attention in detailed analysis, makes significant contributions both individually and working as a team.

**Technical Skills:**

| Programming Languages | Python, Scala, SQL, Java, C/C++, Shell Scripting |
| --- | --- |
| Python Libraries | Pandas, Scikit-Learn, Tensorflow, Pytorch, Matplotlib |
| Hadoop Technologies | Hive, Spark, HDFS, Map Reduce, YARN, Kafka, Pig, Hive, Sqoop, HBase, Oozie, Zookeeper, Cloudera Manager and Horton Works |
| Cloud Platforms | Azure Cloud, AWS Cloud architecture, AWS Big Data computing, Databricks (Unity Catalog, Delta Lake, Lakehouse Architecture) |
| Machine Learning | Decision Tree, SVM, Naïve Bayes, Regression, Classification, Neutral networks |
| Databases | Oracle, DB2, MySQL, Sybase, PostgreSQL. MS SQL Server |
| NoSQL Databases | MongoDB, Cassandra, HBase |
| Visualization Tools | Tableau, Power BI, Advanced Excel |
| Version Control Systems | Git, GitHub, CVS, SVN |
| IDEs | PyCharm, IntelliJ IDEA, Jupyter Notebooks, Eclipse |
| Operating Systems | Unix, Linux, Windows |

**Professional Experience:**

**Client: Fifth Third Bank - Cincinnati, Ohio || December 2023 - Present**

**Sr. Data Engineer**

**Responsibilities:**

* Participated in requirement grooming meetings which involves understanding functional requirements from a business perspective and providing estimates to convert those requirements into software solutions.
* Performed analysis on existing data flows and created high level/low level technical design documents for business stakeholders that confirm technical design aligns with business requirements.
* Used Pandas, Scikit-Learn, Tensorflow, Pytorch, Matplotlib in Python for developing data pipelines and various machine learning algorithms.
* Designed and developed **ETL pipelines** for processing large-scale data using **Apache Spark, Hadoop, and PySpark.**
* Implemented Oracle PL/SQL procedures and packages for ETL processes, improving data ingestion performance by 30%.
* Designed and implemented Kafka-based real-time data pipelines for fraud detection, reducing alert latency to under 300 ms, improving response time for the security operations team.
* Designed and developed scalable data pipelines using Cloud Dataflow, Cloud Composer (Apache Airflow), and Cloud Pub/Sub for real-time and batch data processing.
* Led database tuning efforts, including index optimization, query tuning, and partitioning strategies in Oracle environments.
* Standardized data classification and governance policies using Unity Catalog’s managed catalogs, schemas, and tables across dev, QA, and prod environments.
* Worked on creation and deployment of Spark jobs in different environments and loading data to NoSQL database Cassandra/Hive/HDFS. Secure the data by implementing encryption-based.
* Developed code using: Apache Spark and Scala, IntelliJ, NoSQL databases (Cassandra), Jenkins, Docker pipelines, GITHUB, Kubernetes, HDFS file System, Hive, Kafka for streaming Real time streaming data, Kibana for monitor logs etc. authentication/authorization to the data.
* Integrated Kafka Streams and Apache Spark Structured Streaming to support scalable ingestion and transformation of high-velocity event data.
* Implemented data ingestion from external APIs and cloud sources into GCP BigQuery via Cloud Functions and Cloud Dataflow, supporting near real-time analytics.
* Ensured multi-tenant data isolation and access controls in shared lakehouse architecture by implementing secure view patterns with Unity Catalog.
* Designed end-to-end data ingestion and transformation flows using SnapLogic for enterprise-level data warehousing and analytics initiatives.
* Participated in early adoption and optimization of Unity Catalog in Databricks SQL to enable governed self-service BI.
* Led the migration of on-premise data infrastructure to GCP, optimizing for scalability, performance, and cost using services like BigQuery, Cloud Storage, and Cloud SQL.
* Extracted, Transformed and Loaded data from Sources Systems to Azure Data Storage services using a combination of Azure Data Factory, T-SQL, Spark SQL, and U-SQL Azure Data Lake Analytics.
* Worked on Data Ingestion to one or more Azure Services - (Azure Data Lake, Azure Storage, Azure SQL, and Azure DW) and processing the data in Azure Databricks.
* Ingested API data into Azure Data Storage services (Blob) using Eventhub and PySpark in Databricks. Data was further transformed into a structured format and stored in DataLake gen2.
* Implemented business logic to perform complex data operations, created Dim and Fact tables in Data warehouse using Data factory as orchestration tool, developed a layer SSAS cube which in-turn was used by PowerBI/Tableau for reporting dashboard services, created automated logic app workflows to send out daily reports.
* Leveraged SnapLogic’s API Management features to expose and consume REST APIs for real-time application integrations.
* Created and optimized dbt macros, sources, and models to adhere to best practices and maintain data quality across analytics layers.
* Developed and maintained Oracle Data Warehouses, handling datasets exceeding 10TB, ensuring high availability and consistency.
* Led the migration of on-premise data infrastructure to GCP, optimizing for scalability, performance, and cost using services like BigQuery, Cloud Storage, and Cloud SQL.
* Worked on PySpark and Spark SQL transformation in Azure Databricks to perform complex transformations for business rule implementation.
* Integrated dbt with Git for version control and implemented CI/CD pipelines for automated deployments.
* Optimized **SQL and NoSQL queries for** faster data retrieval in **BigQuery, MongoDB, and Amazon Redshift.**
* Migrated data pipelines, from azure to snowflake multi-cluster data warehouse as it handles semi-structured data which are in turn connected to analytical engines for reporting.
* Built deployment automation ETL CI/CD pipelines for deployment of API solutions and database changes using, Azure DevOps, maintained codebase in azure git, wrote unit/regression/integration tests for successful production deployment and post-deployment validations.
* Created several Databricks Sparkjobs with PySpark to perform several tables to table operations
* Used Apache Airflow to build data pipelines and uses various airflow operators like bash operator and Hadoop operator and branching operator.
* Configured and implemented the Azure Data Factory Triggers and scheduled the Pipelines. Created and configured a new event hub with the provided event hubs namespace.
* Responsible for deployments to DEV, QA, PRE-PROD (CERT) and PROD using Azure.
* Responsible for facilitating load data pipelines and benchmarking the developed product with the set performance standards.
* Used SnapLogic Data Catalog and Manager for organizing, scheduling, and monitoring pipelines efficiently.
* Developed and produced a dashboard, key performance indicators and monitor organization performance.
* Defined data needs, evaluated data quality, and extracted/transformed data for analytic projects and research.
* Tested the database using complex SQL scripts and handling the performance issues effectively.
* Worked with business users to flush out requirements, envision what data (KPI) is required and how it will look after its pumped through Tableau & Power BI tools from Amazon kinesis on real-time basis.

**Environment:** Python, Pandas, Scikit-Learn, Tensorflow, Pytorch, Matplotlib, Azure Data Lake, Azure Storage, Azure Data Factory, Azure SQL, Azure DW, Azure DevOps, Databricks, Snowflake Schema, Machine Learning, ETL, Hadoop, Spark, Oracle, PySpark, Kafka, REST API, PySpark, SQL, Power BI, Apache Airflow, Jenkins, Power BI, Tableau

**Client: Travelers - New York City, New York || November 2020 - November 2023**

**Sr. Data Engineer**

**Responsibilities:**

* Built machine learning predictive ETL pipelines which was accessed using REST API built out of geological data, interpreting seismic data and for optimizing oil exploration process.
* Developed Spark applications using PySpark and Spark-SQL for data extraction, transformation and aggregation from multiple file formats for analyzing and transforming the data to uncover insights into the customer usage patterns.
* Used GCP Data Loss Prevention (DLP) and IAM to implement enterprise-grade data security, compliance, and access control policies across the platform.
* Designed and implemented scalable data warehouse solutions using Snowflake’s virtual warehouse and multi‑cluster architecture.
* Built environment and deployment automation, infrastructure-as-code, deployment data pipeline specification and development using utilities like GitHub Actions using Terraform, Azure DevOps and CI/CD automated pipelines for release cycles.
* Implemented Azure Logic Apps, Azure Functions, Azure Storage, and Service Bus queries for large enterprise level ERP integration systems.
* Designed and maintained ETL/ELT pipelines in Palantir Foundry, integrating data from cloud platforms (AWS S3, Azure Blob), APIs, and on-prem databases.
* Built complex ETL jobs that transform data visually with data flows or by using compute services Azure Databricks, and Azure SQL Database.
* Automated data extraction, transformation, and loading (ETL) workflows using Oracle Data Integrator (ODI) and SQL Loader.
* Implemented Unity Catalog across multiple workspaces in Databricks to centrally manage metadata, access controls, and lineage across teams and projects.
* Integrated Palantir Foundry with DevOps CI/CD pipelines, ensuring automated deployment, version control, and continuous integration.
* Worked in Azure environment for development and deployment of Custom Hadoop Applications.
* Designed and implemented scalable Cloud Data and Analytical architecture solutions for various public and private cloud platforms using Azure.
* Led the migration of on-premise data infrastructure to GCP, optimizing for scalability, performance, and cost using services like BigQuery, Cloud Storage, and Cloud SQL.
* Implemented various Azure platforms such as Azure SQL Database, Azure SQL Data Warehouse, Azure Analysis Services, HDInsight, Azure Data Lake and Data Factory.
* Extracted and loaded data into Data Lake environment (MS Azure) by using Sqoop which was accessed by business users.
* Integrated Oracle databases with AWS/GCP BigQuery/Snowflake environments for hybrid cloud solutions.
* Built end-to-end data pipelines on GCP using Cloud Dataflow, Cloud Pub/Sub, and BigQuery to process real-time and batch data with high throughput and low latency.
* Migrated data warehouses to Snowflake Data warehouse. Defined virtual warehouse sizing for Snowflake for different type of workloads.
* Led the migration of on-premise data infrastructure to GCP, optimizing for scalability, performance, and cost using services like BigQuery, Cloud Storage, and Cloud SQL.
* Extracted data from Data Lakes, EDW to relational databases for analyzing and getting more meaningful insights using SQL Queries and PySpark.
* Worked on Spark improving the performance and optimization of the existing algorithms in Hadoop using Spark Context, Spark-SQL, Data Frame, Pair RDD's
* Utilized Spark-SQL & PySpark based models on Azure Databricks ML workspace, to build & deploy ML models at scale for suspicious phishing, covid-19 impact, customer activity models, ecosystem, retention/stickiness etc.
* Developed PySpark script to merge static and dynamic files and cleanse the data.
* Designed, developed and did maintenance of data integration programs in a Hadoop and RDBMS environment with both traditional and non-traditional source systems.
* Created PySpark procedures, functions, packages to load data.
* Developed MapReduce programs to parse the raw data, populate staging tables and store the refined data in partitioned tables in the EDW.
* Worked on monitoring and troubleshooting the Kafka-storm-HDFS data pipeline for real time data ingestion in Data Lake in HDFS. Wrote Sqoop Scripts for importing and exporting data from RDBMS to HDFS.
* Solved performance issues in Hive and Pig scripts with understanding of Joins, Group and Aggregation and how it translates to MapReduce jobs.
* Designed workflows using Airflow to automate the services developed for change data capture.
* Carried out data transformation and cleansing using SQL queries and PySpark.
* Used Kafka and Spark streaming to ingest real time or near real time data in HDFS.
* Worked on PySpark APIs for data transformations.
* Created dashboards as part of Data Visualization using Power BI.

**Environment:** Python, Azure Cloud (Databricks, Data Factory, Cosmos DB, Function App, Data Lake Gen2, Logic Apps, Azure Storage, Service Bus, Azure SQL), Oracle, Machine Learning, Hadoop, Spark, Spark SQL, PySpark, Hive, HDFS, Pig, Sqoop, MapReduce, Apache Kafka, Airflow, Snowflake, SQL DB, SQL, Power BI, Git

**Client: Walgreens Boots Alliance - Deerfield, Illinois || March 2018 - October 2020**

**Data Engineer**

**Responsibilities:**

* Built high-quality Data warehouses and data lakes at enterprise level, worked with cross functional teams to automate data ingestion and schedule jobs at daily, weekly & monthly frequencies on AWS cloud.
* Modernized the data warehouse environment by using AWS cloud-based Hadoop system, data manipulations, and maintained tables for reporting, data science, and dash boarding and ad-hoc analyses.
* Utilized Python, Spark on AWS Elastic Search, Sagemaker to develop & execute Analytics & Machine learning models for fraud detection and risks.
* Designed and developed scalable data pipelines using Cloud Dataflow, Cloud Composer (Apache Airflow), and Cloud Pub/Sub for real-time and batch data processing.
* Developed ETL Processes in AWS Glue to migrate data from external sources like S3, ORC/Parquet/Text Files into AWS Redshift.
* Used GCP Data Loss Prevention (DLP) and IAM to implement enterprise-grade data security, compliance, and access control policies across the platform.
* Led migrations from legacy data platforms to Palantir Foundry, ensuring minimal downtime and data integrity.
* Conducted performance tuning and troubleshooting for Foundry Data Pipelines, enhancing scalability and efficiency.
* Developed and scaled AWS Lambda + Kinesis pipelines to process over 5 TB of real-time streaming data per day, enabling near-instant insights for operational dashboards.
* Optimized real-time ETL workflows using AWS MSK, improving throughput by 40% and ensuring high availability with autoscaling consumers.
* Designed and developed scalable Snowflake data warehouses, including data modeling (star/snowflake schema), schema design, and performance tuning.
* Built and optimized Snowflake ELT pipelines using Snowpipe, Streams, Tasks, and external stages (S3/GCS) to load real-time and batch data.
* Implemented role-based access control (RBAC), masking policies, and data sharing features to ensure security and collaboration across business units.
* Created AWS Lambda functions and assigned IAM roles to schedule python scripts using Cloud Watch Triggers to support the infrastructure needs (SQS, Event Bridge, SNS)
* Built logical and physical data models for snowflake for different types of workloads, developed large scale data intelligence solutions around data warehouse.
* Ingested data by going through cleansing and transformations and leveraging AWS Lambda, AWS Glue and Step Functions.
* Used Python for pattern matching in build logs to format warnings and errors.
* Designed and deployed Hadoop cluster and different big data analytic tools including Pig, Hive, HBase and Sqoop.
* Developed simple and complex MapReduce programs in Hive, Pig and Python for Data Analysis on different data formats.
* Designed, developed, and maintained scalable data pipelines using Palantir Foundry Transform Pipelines, processing terabytes of data daily.
* Wrote Spark applications using Scala to interact with the database using Spark SQL Context and accessed Hive tables using Hive Context.
* Involved in designing different components of system like big-data event processing framework Spark, distributed messaging system Kafka and SQL database.
* Used PySpark data frame to read text data, CSV data, and image data from HDFS, S3 and Hive.
* Managed and scheduled Jobs on Hadoop Cluster using UC4 (Confidential Preoperatory Scheduling Tool) workflows. Managed and reviewed Hadoop and HBase log files.
* Created monitoring dashboards and alerts with GCP Stackdriver (Cloud Monitoring and Logging) to proactively manage data pipeline health and system performance.
* Imported semi-structured data from Avro files using Pig to make serialization faster
* Managed large datasets using Panda data frames and MySQL.
* Implemented Spark Streaming and Spark SQL using Data Frames.
* Created multiple Hive tables, implemented Dynamic Partitioning and Buckets in Hive for efficient data access.
* Involved in converting MapReduce programs into Spark transformations using Spark RDD on Python.
* Implemented data quality, integrity and reliability and creating spark performance tuning solutions through the data pipeline by designing, maintaining, and promoting data governance, enrichment & database efficiency.
* Extracted data from medical devices stored by scientists in different network drives, loaded into NOSQL DB (Mongo) to expedite research studies from internal & third-party sources for the R&D department through use of algorithms.
* Involved in building unit tests and integration tests for automated data validation quality checks before deployment.
* Designed and developed scalable Snowflake data warehouses, including data modeling (star/snowflake schema), schema design, and performance tuning.
* Built and optimized Snowflake ELT pipelines using Snowpipe, Streams, Tasks, and external stages (S3/GCS) to load real-time and batch data.
* Implemented role-based access control (RBAC), masking policies, and data sharing features to ensure security and collaboration across business units.
* Built the analytics data warehouse and cubes on cloud- based Hadoop, Git version control system, automatic deployment using Jenkins & scheduling system.
* Created several complex Reports (sub reports, graphical, multiple groupings, drilldowns, parameter driven, formulas, summarized and conditional formatting) in Power BI for visualization of data.

**Environment:** Python, AWS (S3, Elastic Search, Sagemaker, Glue, Redshift, SQS, SNS, Lambda, Step Functions, Oracle, Event Bridge, CloudWatch, IAM), Machine Learning, Snowflake, Hadoop, Spark, Spark SQL, Scala, HBase, Hive, Pig, Sqoop, MapReduce, HDFS, Kafka, MongoDB, Git, Jenkins, Power BI

**Client: Tally Solutions - Bengaluru, India || November 2016 - December 2017**

**Hadoop Data Engineer**

**Responsibilities:**

* Responsible for continuous monitoring and managing the Hadoop cluster through Hortonworks (HDP) distribution.
* Worked on Spark RDD, Data Frame API, Data set API, Data Source API, Spark SQL and Spark Streaming.
* Developed Spark Applications by using Python and implemented Apache Spark data processing Project to handle data from various RDBMS and Streaming sources.
* Worked with the Spark for improving performance and optimization of the existing algorithms in Hadoop.
* Used Spark Context, Spark-SQL, Spark MLlib, Data Frame, Pair RDD and Spark YARN.
* Used Spark Streaming APIs to perform transformations and actions on the fly for building common.
* Developed Kafka consumer API in python for consuming data from Kafka topics.
* Consumed Extensible Markup Language (XML) messages using Kafka and processed the XML file using Spark Streaming to capture User Interface (UI) updates.
* Developed Pre-processing job using Spark Data frames to flatten JSON documents to flat file.
* Wrote live Real-time Processing and core jobs using Spark Streaming with Kafka as a Data pipeline system.
* Worked on AWS Data Pipeline to configure data loads from S3 to into Redshift and have used AWS components (Amazon Web Services) - Downloading and uploading data files (with ETL) to AWS system using S3 components.
* Used AWS Data Pipeline to schedule an Amazon EMR cluster to clean and process web server logs stored in Amazon S3 bucket.
* Developed Lambda functions and assigned IAM roles to run Python scripts along with various triggers (SQS, SNS).
* Developed and deployed AWS Lambda services for ETL migration services by generating a serverless data pipeline which can be written to Glue and queried from Athena.
* Stored incoming data in the Snowflakes staging area. Created numerous ODI interfaces and load into Snowflake DB.
* Designed columnar families in Cassandra and ingested data from RDBMS, performed data transformations, and then exported the transformed data to Cassandra as per the business requirement.
* Experienced in creating data-models for clients transactional logs, analyzed the data from Cassandra.
* Used Hive QL to analyze the partitioned and bucketed data, Executed Hive queries on Parquet tables.
* Stored in Hive to perform data analysis to meet the business specification logic.
* Developed Custom UDF in Python and used UDFs for sorting and preparing the data.
* Worked on Custom Loaders and Storage Classes in PIG to work on several data formats like JSON, XML, CSV and generated Bags for processing using Pig etc.
* Developed Sqoop and Kafka Jobs to load data from RDBMS, External Systems into HDFS and HIVE.
* Developed Oozie coordinators to schedule Hive scripts to create Data pipelines.
* Designed and developed scalable data pipelines using Cloud Dataflow, Cloud Composer (Apache Airflow), and Cloud Pub/Sub for real-time and batch data processing.
* Wrote several Map Reduce Jobs using PySpark, NumPy and used Jenkins for Continuous integration.
* Worked on cluster and testing of HDFS, Hive, Pig and MapReduce to access cluster for new users.
* Responsible for continuous monitoring and managing the Hadoop cluster through Cloudera Manager.
* Maintained Tableau functional reports based on user requirements.

**Environment:** Python, Amazon Web Services (AWS), Spark, Spark-Streaming, Spark SQL, MapReduce, HDFS, Hive, Pig, Apache Kafka, Sqoop, Python, PySpark, Shell scripting, Linux, MySQL, Oracle Enterprise DB, Jenkins, Eclipse, Oracle, Git, Oozie, Tableau, MySQL, Soap, XML, Cassandra, Agile Methodologies

**Client: Larsen & Toubro Infotech (LTI) - Mumbai, India || July 2014 - October 2016**

**Big Data Engineer**

**Responsibilities:**

* Worked with Hadoop Ecosystems components like HBase, Sqoop, Zookeeper, Oozie, Hive and Pig with Cloudera Hadoop distribution.
* Involved in start to end process of Hadoop jobs that used various technologies such as Sqoop, Pig, Hive, MapReduce, Spark, and Shells scripts.
* Managed and supported enterprise Data Warehouse operation, Big Data application development using Cloudera and Hortonworks HDP.
* Developed Pig scripts to transform the raw data into intelligent data as specified by business users.
* Installed Hadoop, Map Reduce, HDFS to develop multiple MapReduce jobs in Pig and Hive for data cleansing and pre-processing.
* Used Spark API over Hortonworks Hadoop YARN to perform analytics on data in Hive.
* Improved the performance and optimization of the existing algorithms in Hadoop using SparkContext, Spark-SQL, Data Frame, Pair RDD's, Spark YARN.
* Developed Spark code using Scala and Spark-SQL/Streaming for faster testing and processing of data.
* Developed a Spark job in Java which indexes data into Elastic Search from external Hive tables which are in HDFS.
* Performed transformations, cleaning and filtering on imported data using Hive, MapReduce, and loaded final data into HDFS.
* Explored with the Spark improving the performance and optimization of the existing algorithms in Hadoop using Spark Context, Spark-SQL, Data Frame, Pair RDD's, Spark YARN.
* Import the data from different sources like HDFS/HBase into Spark RDD and developed a data pipeline using Kafka and Storm to store data into HDFS.
* Used Spark streaming to receive real time data from the Kafka and store the stream data to HDFS using Scala and NoSQL databases such as HBase and Cassandra.
* Documented the requirements including the available code which should be implemented using Spark, Hive, HDFS and HBase.
* Managed GCP IAM roles, security policies, and networking components (VPC, Firewall Rules) to ensure secure and compliant data pipelines.
* Utilized PySpark to distribute data processing on large streaming datasets to improve ingestion and process.
* Performed transformations like event joins, filter boot traffic and some pre-aggregations using Pig.
* Executed Hive queries on Parquet tables stored in Hive to perform data analysis to meet the business requirements.
* Configured Oozie workflow to run multiple Hive and Pig jobs which run independently with time and data availability.
* Imported and exported the analyzed data to the relational databases using Sqoop for visualization and to generate reports for the BI team.
* Created several types of data visualizations using Python and Tableau.

**Environment:** Cloudera, Hortonworks, Hadoop, Sqoop, Pig, Hive, MapReduce, Spark, PySpark, Shell Scripting, SQL, Hortonworks, Python, Java, Scala, MLLib, HDFS, YARN, Kafka, Zookeeper, Oozie, Cassandra, Tableau, Agile