**Sateesh Kantamani**

**Pavanasateesh2729@gmail.com**

**208-991-2729**

**Professional Summary:**

Over 15 years experience in the IT industry, including 9 years of experience as a Data Engineer and Data Analyst. I bring a comprehensive skill set that covers various aspects of data engineering, data analysis, cloud platform expertise and software testing. My proficiency spans both AWS and Azure environments, allowing me to contribute effectively to diverse projects.

**Background includes:**

* Strong understanding of Data Modelling, Data Validation, Data Source Evaluation, **ETL**, BI, and OLAP, as well as client/server applications on **AWS** and **Azure**.
* Expertise in configuring servers for auto-scaling and elastic load balancing in **AWS** services such as EMR, Redshift, S3, and EC2.
* Experience using **AWS** S3 for staging, transferring, and archiving data, alongside **AWS** Redshift for large-scale data migrations, change data collection, and change data management in **AWS** DMS.
* Expertise in **Python** scripting, complex **SQL** query crafting, and query optimization for Oracle, **SQL** Server, and Teradata. A solid understanding of the Software Development Life Cycle (SDLC) and testing methodologies.
* Excellent Data Analysis, Data Validation, Data Cleansing, and Data Verification skills, along with the ability to identify data consistency issues.
* Expertise with complex **SQL** across Oracle and Teradata sources for data analysis and profiling.
* Proficient in **Azure** Data Factory, **T-SQL**, and **Spark** **SQL** for extracting, transforming, and loading data.
* Experience with big data technologies including **Hadoop**, **Spark**, **Python**, **SQL**, Power BI, Tableau, and other tools used for Data Engineering on both **Azure** and **AWS** cloud platforms.
* Thorough understanding of big data technologies, dashboard design, **ETL** pipeline construction, and data warehousing.
* Expertise in processing data in **Azure** Databricks and **Azure** Data Factory (ADFv2) after specializing in data intake to **Azure** services like **Azure** Data Lake, **Azure** storage, and **Azure** **SQL**.
* Practical knowledge in utilizing a range of **Azure** Cloud technologies, such as **Azure** **SQL** Database, **Azure** Data Factory, **Azure** Databricks, **Azure** Data Lake Store, **Azure** Synapse Analytics, **Azure** Storage Accounts (Gen1 & Gen2), **Azure** Key Vault, and Smart Home Hub.
* Competence in the use of several **ETL** tools, including Informatica Power Center, for data migration, profiling, ingestion, cleansing, transformation, import, and export.
* Practical knowledge of **Spark** RDD, **Spark** **SQL**, **Spark** Streaming, Data Frame API, Data Set API, and Data Source API.
* Expertise with Hive query performance optimization and data export/import using Sqoop between HDFS and relational database platforms.
* Expert in using Teradata, Amazon Redshift, **Azure** Data Warehouse, Business Objects, SSAS/SSRS/SSIS, and other business intelligence and data warehouse tools.
* Expertise in developing real-time data streaming solutions with Flume, **Kafka**, and Apache **Spark**/**Spark** Streaming.
* In charge of finding and putting into practice new **SQL** Server features that improve query performance through in-memory optimization and **T-SQL** language advancements.
* Managed messaging services with Apache **Kafka** and automated data transfer between various **Hadoop** systems using Apache NiFi.
* Capable of utilizing **ETL** tools for data warehousing, creating reports and dashboards in Tableau, and mining data.
* Efficient cooperation inside the cloud ecosystems of **GCP** and **Azure**, guaranteeing smooth integration and coherence.
* I have vast experience working on IT data analytics projects, including direct experience moving on-premises **ETL**s to Google Cloud Platform (**GCP**) by hand utilizing native tools like Composer, Cloud Data Proc, BIG query, and Google Cloud Storage.
* Thorough familiarity with job workflow scheduling and locking services/tools such as Zookeeper and Oozie.
* Solid knowledge of No**SQL** databases, including **Cassandra** and HBase.
* Well-versed with Amazon Web Services (**AWS**), including Lambda, Redshift, DynamoDB, EMR, Step functions, and EC2 for processing and storage.
* Practical knowledge of Microsoft **Azure** services such as Logic Apps, Data Factory, BLOB, ADLS, and HDInsight Clusters.

**Technical Skills:**

|  |  |
| --- | --- |
| Hadoop Components / Big Data | PySpark, Airflow, HDFS, MapReduce, Hive, HCatalog, HBase, Sqoop, Impala, Zookeeper, Kafka, Yarn. |
| Programming Languages | Python, Scala, SQL, PySpark, PowerShell, T-SQL. |
| Cloud Platform | AWS (Lambda, S3, EC2, EMR, RDS), Microsoft Azure (Azure Databricks, Azure Data Factory, Azure Data Explorer, Azure HDInsight, ADLS), GCP |
| Reporting and ETL Tools | AWS GLUE, Tableau, Power BI. |
| Databases | Oracle, SQL Server, MS Access, NoSQL Database (HBase, MongoDB, DynamoDB) |
| Big Data Technologies | Hadoop, HDFS, Hive, Oozie, Sqoop, Spark, Machine Learning, Pandas, NumPy, Seaborn, Impala, Zookeeper, Airflow, Informatica, Snowflake, Data Bricks, Kafka, Cloudera |
| Data Analysis Libraries: | Pandas, NumPy, SciPy, Scikit-learn, Matplotlib |
| Containerization | Docker, Kubernetes |
| CI/CD Tools | Jenkins, GitLab, Bamboo. |
| Software Methodologies | Agile, Scrum, Waterfall |
| Development Tools | Eclipse, PyCharm, IntelliJ, SSMS, Microsoft Office Suite (Word, Excel, PowerPoint) |
| Version Control Tools | GitHub and Azure DevOps |
| J2SE/J2EE Technologies | Java, J2EE, RMI, Sockets, JDBC, Servlets, JSP, JMS, Java Beans, JSTL, Jakarta Struts, JSF, EJB, Spring, Hibernate, JTA, JNDI, JPA, JMS. |

**Educational Qualification:**

* MBA from SV University, Tirupati 2012, India.
* B.com computer science Andhra University 2003.

**Professional Experience:**

**Client: USAA, San Antonio, TX Mar 2023 – Till date**

**Role: Sr. Data Engineer:**

**Responsibilities:**

* Collaborated closely with business/user groups to elicit requirements and played a key role in the creation and development of pipelines.
* Utilized Airflow for building batch data pipelines, while **Kafka** was employed for constructing streaming data pipelines.
* Created a **Spark** Streaming script that consumes topics from **Kafka** for real-time processing, enhancing data accuracy.
* Developed and managed **Azure** Data Factory, incorporating Blob storage for storage and backup on the **Azure** platform.
* Leveraged JavaScript, HTML5, CSS3, Bootstrap, Ajax, JSON, XML, jQuery, Angular.js, React.js, and Node.js to design interactive front-end interfaces.
* Applied **Spark** applications in **Python** to establish distributed environments for loading high-volume files using **PySpark** into **Azure** **SQL** DB tables.
* Wrote Java API for Amazon Lambda to manage specific **AWS** services.
* Designed and automated pipelines using Databricks for **ETL** processes, ensuring efficient workflow management.
* Engineered **ETL** packages in SSIS to extract and maintain data from various sources, such as Access databases, Excel Spreadsheets, and flat files.
* Operated **ETL** processes using **Azure** Databricks, employing **Kafka** for connecting to relational databases and Informatica for session and workflow creation and monitoring.
* Automated data ingestion, transformation, and storage using Apache **Spark** and Delta Lake, ensuring data quality and integrity.
* Developed data workflows utilizing Data bricks, Scala, and **Spark**, capturing data from Delta tables in Delta Lakes.
* Implemented streaming pipelines via **Azure** Event Hubs and Stream Analytics, facilitating data-driven workflow analysis.
* Utilized Delta Lakes for consistent data unification and worked on ACID transactions using Apache **Spark** for data processing.
* Took charge of identifying and implementing **SQL** Server enhancements to optimize query performance and ensure data integrity.
* Implement security best practices and ensure compliance with data governance standards while handling sensitive data within AWS and Snowflake environments.
* Served as a subject matter expert on performance tuning stored procedures, functions, **T-SQL** scripts, indexes, and SSIS packages.
* Managed data cleansing and performance tuning tasks through various SSIS transformations.
* Performed application deployment and SOA middleware performance tuning in clustered SOA 12c environments.
* Implemented a distributed stream processing platform for low-latency data integration inside and outside **Azure**.
* Utilized PowerShell scripting for data maintenance and configuration, while Apache Airflow automated and validated data processes.
* Integrate data from diverse sources, both internal and external, using AWS services like Data Pipeline, Lambda, or other suitable methods, ensuring seamless data flow.
* Enhanced Hive queries using best practices, **Hadoop**, YARN, **Python**, and **PySpark** to optimize query performance.
* Deployed Sqoop for data extraction from Teradata to HDFS and the subsequent analysis of patterns back to Teradata.
* Employed **Kafka** to collect data from sources and store it in HDFS systems for further filtering.
* Applied Accumulators and Broadcast variables to monitor and optimize **Spark** applications and analytics.
* Oversaw **Hadoop** cluster job performance and capacity planning, ensuring high availability and cluster recovery.
* Participated in a Data Migration project, transferring data from different sources to Google Cloud Platform using **Spark** Scala scripts.
* Designed data ingestion processes for maintaining a Global Data Lake on the **GCP** cloud and Big Query.
* Created Tableau reports and dashboards, tailored to business requirements.
* Followed Agile methodology throughout all phases of the Software Development Life Cycle.

**Environment**: Python, SQL, Cassandra DB, Azure Data Factory, Azure SQL DB, Spark, Databricks, SSIS, SQL Server, Kafka, Informatica, Apache Spark, Delta Lake, Azure Event Hubs, Stream Analytics, Azure Blob Storage, PowerShell, Apache Airflow, Hadoop, YARN, PySpark, Hive, Teradata, Sqoop, HDFS, Spark, Agile.

**Client: WebMD LLC, New York, NY Oct 2020 – Jan 2023**

**Role: Sr. Data Engineer**

**Responsibilities:**

* Contributed to the entire Software Development Life Cycle (SDLC) by analyzing business requirements and comprehending the functional flow of information from source to destination systems.
* Leveraged analytical, statistical, and programming skills to gather, analyze, and interpret extensive datasets for the creation of data-driven solutions using tools like **SQL** and **Python**.
* Designed **AWS** EC2 instance architecture in alignment with high availability application structures and security criteria.
* Established **AWS** S3 buckets, managed associated policies, and harnessed S3 and Glacier for storage and backup. Utilized **Hadoop** clusters and data querying tools for efficient data storage and retrieval.
* Created and implemented SSIS Packages to facilitate data import and export from sources such as MS Excel, **SQL** Server, and Flat files.
* Engaged in the design, development, and testing of J2EE components encompassing Java Beans, Java, XML, Collection Framework, JSP, Servlets, and JDBC.
* Developed UI pages using HTML5, Bootstrap, CSS3, JavaScript, jQuery, AJAX, React.js, and Node.js.
* Executed Data Integration tasks including data extraction, transformation, and loading processes within designed packages.
* Deployed automated **ETL** workflows through **AWS** Lambda, organized and cleansed data in S3 buckets using **AWS** Glue, and executed data processing in Amazon Redshift.
* Contributed to **ETL** architecture enhancements to optimize performance through query optimization.
* Utilized **Spark** and Hive for processing data extracted from large datasets in HDFS.
* Handled Streaming data transfers from diverse sources into HDFS and No**SQL** databases.
* Designed **ETL** Mappings in Talend Integration Suite for extracting, transforming, and loading data from sources into target databases.
* Performed **SQL** query optimization to enhance performance using query analyzer.
* Utilized **Python** scripting in **Spark** for data transformation from various file formats, including Text files, CSV, and JSON.
* Imported data from different relational databases like My**SQL** and Teradata using Sqoop jobs.
* Utilized **Python** for data processing and testing through **Spark** **SQL**, real-time processing with **Spark** Streaming and **Kafka**.
* Scripted in **Python** and PowerShell to establish baselines, manage branching, merging, and automation processes with GIT.
* Enhanced **ETL** architecture and optimized workflows by constructing DAGs in Apache Airflow for scheduling **ETL** jobs. Implemented additional Apache Airflow components like Pools, Executors, and multi-node functionality.
* Implemented various transformations in SSIS Dataflow and Control Flow, including for loop Containers and Fuzzy logic.
* Created SSIS packages for Data Conversion, employing data conversion transformations, and generating advanced extensible reports using **SQL** Server Reporting Services.
* Deployed applications on **GCP** using Spinnaker (rpm based), initiated multi-node Kubernetes clusters in Google Kubernetes Engine (GKE), and migrated dockerized applications from **AWS** to **GCP**.

**Environment:** Python, SQL, AWS EC2, AWS S3 buckets, Hadoop, PySpark, AWS Lambda, AWS Glue, Amazon Redshift, Spark Streaming, GCP, Apache Kafka, SSIS, Informatica, ETL, Hive, HDFS, NoSQL, Talend, MySQL, Teradata, Sqoop, PowerShell, GIT, Apache Airflow.

**Client: ETCC – Plano, TX Apr 2018 – July 2020**

**Data Engineer**

**Responsibilities:**

* Involved in open-source Apache Distribution and **Hadoop** administration, manual configuration of Core-Site, HDFS-Site, YARN-Site, and Map Red-Site is required. However, with widely adopted **Hadoop** distributions like Hortonworks, Cloudera, or MapR, configuration files are automatically set up during startup, eliminating the need for manual configuration by **Hadoop** administrators.
* Utilized Sqoop for importing data from Relational Databases such as My**SQL** and Oracle.
* Engaged in the import of both structured and unstructured data into HDFS.
* Accountable for fetching real-time data through **Kafka** and executing processing using **Spark** and Scala.
* Leveraged **Kafka** to import real-time weblogs and subsequently ingested the data into **Spark** Streaming.
* Developed business logic using **Kafka** Direct Stream within **Spark** Streaming, implementing necessary business transformations.
* Utilize AWS Lambda to process and analyze real-time data streams, enabling timely insights and actions based on dynamic data sources.
* Created and executed a real-time streaming **ETL** pipeline utilizing **Kafka** Streams API.
* Utilized Hive to implement Web Interfacing and store data within Hive tables.
* Migrated Map Reduce programs to **Spark** transformations using **Spark** and Scala.
* Possess familiarity with **Spark** Context, **Spark**-**SQL**, and **Spark** YARN.
* Handled data extraction and ingestion from various sources into the **Hadoop** Data Lake by constructing **ETL** pipelines via Sqoop, Hive, and **Spark**.
* Took charge of data importation to HDFS via Sqoop from different RDBMS servers, along with exporting aggregated data using Sqoop back to the RDBMS servers for additional **ETL** operations.
* Designed and developed **PySpark** applications using **Python** to evaluate **Spark**'s performance against Hive.
* Created **Spark** modules, crafted intricate queries, functions, and views to generate reports for business users.
* Generated automated **Python** scripts to transform data from diverse sources and create **ETL** pipelines.
* Developed comprehensive end-to-end data processing pipelines, beginning with data reception via distributed messaging systems like **Kafka**, persisting data into Hive, and enhancing data with **PySpark**.
* Implemented strategies like Sequence files, RC files, Map side joins, bucketing, and partitioning in Hive to enhance performance and storage efficiency.
* Wrote Enterprise Java beans (EJBs) to handle Business Logic Processing.
* Designed and constructed user interfaces for an ERP project using Java Server Faces (JSF).
* Developed Java Persistence components, including entity managers and entities, for data persistence.
* Connected the **Cassandra** database to Amazon EMR File System to store the database in S3.
* Implemented Amazon EMR for processing Big Data across a **Hadoop** Cluster on Amazon Elastic Compute Cloud (EC2) and Amazon Simple Storage Service (S3).
* Deployed the project on Amazon EMR with S3 connectivity for backup storage.

**Environment:** Hadoop, Map Reduce, Hive, Spark, Oracle, GitHub, Tableau, UNIX, Cloudera, Kafka, Sqoop, Scala, NIFI, HBase, Amazon EC2, S3.

**Client: T-Systems ICT India Pvt. Ltd. – India Jul 2015 – Apr 2018**

**Role: Data Analyst/Engineer**

**Responsibilities:**

* Design and construct end-to-end data pipelines, utilizing **AWS** services such as Lambda for serverless computing, ensuring efficient and reliable data movement and transformation.
* Leverage Snowflake's cloud-native data warehousing capabilities to design, build, and manage scalable and performant data storage solutions that align with business needs.
* Develop and implement **ETL** processes using **AWS** Glue or custom **Python** scripts to extract, transform, and load data from various sources into Snowflake for further analysis.
* Craft and optimize complex **SQL** queries to retrieve and manipulate data efficiently within Snowflake's data warehouse environment, ensuring optimal query performance.
* Develop **Python** scripts to automate data-related tasks, perform data validations, and execute data transformations, contributing to the efficiency and accuracy of data workflows.
* Implement data quality checks, validations, and monitoring processes to ensure the integrity, consistency, and accuracy of data stored in Snowflake.
* Work closely with cross-functional teams including data analysts, data scientists, and business stakeholders to understand requirements, deliver actionable insights, and support data-driven decision-making.
* Design automated data loading processes that efficiently handle data updates, inserts, and deletes within Snowflake's data warehouse, ensuring data remains up to date.
* Performance Tuning: Continuously monitor and fine-tune data pipelines, **SQL** queries, and **ETL** processes to optimize performance and maintain the responsiveness of analytical workflows.
* Maintain thorough documentation of data pipelines, transformations, and processes, ensuring that colleagues can understand and replicate the data engineering workflows.
* Identify and resolve data-related issues promptly, applying debugging skills and root cause analysis to maintain the reliability of data engineering solutions.
* Analyzed the system for new enhancements/functionalities and performed Impact analysis of the application for implementing **ETL** changes.
* Developed complex ETL pipelines using AWS Glue and Apache Spark, processing millions of records daily with near-real-time processing.
* Built performant, scalable **ETL** processes to load, cleanse and validate data.
* Participated in the full software development lifecycle with requirements, solution design, development, QA implementation, and product support using Scrum and other Agile methodologies.
* Collaborate with team members and stakeholders in design and development of data environment.
* Preparing associated documentation for specifications, requirements, and testing.
* Stay updated with emerging technologies and trends in the data engineering field, incorporating relevant tools and techniques to enhance the overall data ecosystem.

**Environment:** AWS, Snowflake, Lambda, SQL, Python, ETL, Docker, GitHub.

**Client: Bosch Global Software Technologies – India Sept 2012 – Jun 2015**

**Role: ETL Developer**

**Responsibilities:**

* Designed and developed complex Informatica mappings by using Lookup, Expression, Update, Sequence generator, Aggregator, Router, Stored Procedure, etc., transformations to implement complex logics while coding a mapping.
* Written **Python** and bash scripts for scheduling the oracle **ETL** on the Linux based servers.
* Analysis of our source data systems using the IDQ basic profiling features using the ODI developer and the analyst tools before making any **ETL** Designs.
* Created Scenarios and created scheduling for **ETL**, created transformation Logics as part of Pre-**ETL** work.
* Designed and developed custom mappings in ODI as per business requirements.
* Designed, monitored Incremental and full load of data through Data Warehouse.
* Experience in Load Balancing, experience in load balancing agents.
* Implemented incremental logic for stage load mappings and Insert/update logic for all Fact mappings.
* Using ODI Developed, maintained various Packages, Interfaces, Variables, and Models, which populated the Data into the Staging tables and to DWH.
* Extensively used Oracle data Integrator to extract data from legacy (main frames) to Oracle database as per mapping rules.
* Involved in troubleshooting of ODI objects, Optimized Query Performance, improved Workflow Performance and Reliability bases on statistics from logs.
* Handled error through exception handling and configured the event-based scenarios to send the alert emails.
* Extensively developed mappings on various business areas in the project by employing the available Transformations.
* Created/Modified metadata repository (.rpd) as per the data requirements by using OBIEE Admin tool.

**Environment:** Python, Bash scripts, Linux, Oracle Data Integrator (ODI), OBIEE.

**Client: XELA logistics - India Jul 2009 - Sep 2012**

**Role: Manual and Automation Engineer**

**Responsibilities:**

* In Agile environment, attended daily stand-up meetings, reviewed user Stories.
* Analyzed the business requirements documents, created the detailed test design for new functionality.
* Conducted Smoke Testing, Functional Testing, Exploratory Testing, end to end testing, Regression Testing and User Acceptance Testing.
* Performed Front end and Back-end testing of all the Modules using Manual as well as Automation Testing techniques.
* Wrote and executed SQL queries and Joins, Group By clause, having clause etc to validate the data.
* Filed and tracked the defects using JIRA.
* Conducted comprehensive testing of individual software components to identify defects, bugs, and performance bottlenecks, ensuring high-quality deliverables.
* Proficient in creating Micro services Test suite using Postman, and automated Web Services Projects using Rest-Assured.
* Created Selenium automation scripts in Java using TestNG framework and Page Object Model (POM) and configured WebDriver and Maven tool.
* Developed automated tests in partnership with developers for continuous regression testing and reducing manual testing and developed custom libraries for automation using Selenium, Java, Cucumber.
* Followed Agile/Scrum Methodologies for all the project activities.

**Environment:** Manual Testing, Test planning and strategy development, Test case design and execution, Automated testing (Selenium/Java), Defect management (Mantis / JIRA), Collaboration and teamwork, Agile testing methodologies, SQL Server 2012, TestNG, Postman, Rest-Assured.