Sudhansu Sekhar

Email: sudhansusekhar701@gmail.com
Contact: 737-277-2102
LinkedIn: https://www.linkedin.com/in/sudhanshu-s-b396702a7/

Professional Summary

- Possessing over **5 years** of professional experience as a **Data Engineer**, with a specialization in **Big Data, Hadoop Ecosystem, Cloud Engineering, Data Warehousing**.
- Sound Experience with **AWS** services like **Amazon EC2, S3, EMR, Amazon RDS, VPC, Amazon Elastic Load Balancing, IAM, Auto Scaling, Cloud Front, Cloud Watch,** and **Lambda** to trigger resources**.**
- Experience in building data pipelines using **Azure Data Factory**, **Azure Data bricks,** and loading data to **Azure DataLake, Azure SQL Database**, **Azure SQL Data Warehouse to** control and grant database access.
- Good experience with **Azure** services like **HDInsight**, **Stream Analytics**, **Active Directory**, **Blob Storage**, **Cosmos DB**, **Storage Explorer**.
- Strong Hadoop and platform support experience with all the entire suite of tools and services in major **Hadoop Distributions** – **Cloudera**, **Amazon EMR**, **Azure HDInsight**, and **Hortonworks**.
- Proficient in handling and ingesting terabytes of **Streaming data** (Kafka, Spark streaming, Strom), **Batch Data**, **Automation** and **Scheduling** (Oozie, Airflow).
- Profound knowledge in developing production-ready **Spark** applications using Spark Components like **Spark SQL, Data Frames, Datasets, Spark-ML** and **Spark Streaming**.
- Expertise in developing multiple confluent **Kafka** Producers and Consumers to meet business requirements. Store the stream data to HDFS and process it using **Spark**.
- Strong working experience with **SQL** and **NoSQL** databases (**Cosmos DB, MongoDB,** HBase, **Cassandra**), data modeling, tuning, disaster recovery, backup and creating data pipelines.
- Experienced in scripting with **Python** (**PySpark**), **Scala** and **Spark-SQL** for development, aggregation from various file formats such as **XML, JSON, CSV**, **Parquet.**
- Great experience in data analysis using **HiveQL, Hive-ACID** tables, **Pig Latin** queries, custom MapReduce programs and achieved improved performance.
- Extensive knowledge in all phases of **Data Acquisition**, **Data Warehousing** (gathering requirements, design, development, implementation, testing, and documentation), **Data Modeling** (analysis using Star Schema and **Snowflake** for FACT and Dimensions Tables), **Data Processing** and **Data Transformations** (Mapping, Cleansing, Monitoring, Debugging, Performance Tuning and Troubleshooting Hadoop clusters).
- Experience in monitoring document growth and estimating storage size for large **MongoDB** clusters as part of the data life cycle management.
- Hands-on experience on **Ad-hoc** queries, Indexing, Replication, Load balancing, Aggregation in **MongoDB**.
- Expertise in creating **Kubernetes** cluster with cloud formation templates and **PowerShell** scripting to automate deployment in a cloud environment.
- Sound knowledge in developing highly scalable and resilient **Restful APIs, ETL** solutions, and third-party integrations as part of Enterprise Site platform using **Informatica**.
- Experience in using bug tracking and ticketing systems such as **Jira**, and **Remedy**, used **Git** and **SVN** for version control.
- Highly involved in all facets of **SDLC** using **Waterfall** and **Agile Scrum** methodologies.
- Involved in migration of the legacy applications to cloud platform using DevOps tools like **GitHub**, **Jenkins**, **JIRA**, **Docker**, and **Slack**.
- Collaborate with business, production support, engineering team regularly for diving deep on data, effective decision making and to support analytics platforms.

Core Qualifications

Programming Languages:	C, J2EE, SQL, Pig Latin, HiveQL, Scala, Python, Unix Shell Scripting.
Databases:	MS-SQL SERVER, Oracle, MS-Access, MySQL, Teradata, PostgreSQL, DB2.
Big Data Technologies:	HDFS, Yarn, MapReduce, Pig, Hive, HBase, Cassandra, Oozie, Apache Spark, Scala, Impala, Kafka.
Hadoop Distributions:	Apache Hadoop 2.x/1.x, Cloudera CDP, Hortonworks HDP, Amazon EMR (EMR, EC2, EBS, RDS, S3, Glue, Elastic search, Lambda, Kinesis, SQS, Dynamo DB, Redshift, ECS) Azure HDInsight (Data bricks, Data Lake, Blob Storage, Data Factory, SQL DB, SQL DWH, Cosmos DB, Azure DevOps, Active Directory).
NoSQL Database:	Cassandra, MongoDB.
Reporting Tools/ETL Tools:	Informatica, Talend, SSIS, SSRS, SSAS, ER Studio, Tableau, Power BI.
Methodologies:	Agile/Scrum, Waterfall.

Development Tools:	Eclipse, NetBeans, IntelliJ, Hue, Microsoft Office Suite (Word, Excel, PowerPoint, Access)
Operating Systems:	Windows, Macintosh, Linux, Ubuntu, Unix.
Others:	Machine learning, NLP, Stream Sets, Spring Boot, Jupyter Notebook, Docker, Kubernetes, Snowflake, Jenkins, Jira.

Experience

Azure Data Engineer **May 2023 to Present**

Liberty Mutual Inc. Boston, MA

- Managed Azure cloud platforms including **HDInsight, Data bricks, Data Lake, Blob, Data Factory, Synapse, SQLDB, SQL DWH.**
- Performed data cleansing and applied transformations using **Data bricks** and **Spark** data analysis**.**
- Designed and automated Custom-built input adapters using **Spark**, **Sqoop** and **Oozie** to ingest and analyze data from RDBMS to **Azure Data lake.**
- Involved in the development of automated workflows for daily incremental loads, moving data from traditional RDBMSs to data lakes.
- Worked on Azure Synapse analytics service that brings together **enterprise data warehousing and Big Data analytics.**
- Experienced in the creation of database objects such as tables, views, stored procedures, triggers, packages, and functions using **T-SQL** to provide efficient data management and structure.
- Performed Extract Transform and Load data from Sources Systems to Azure Data Storage services using a combination of AzureData Factory, T-SQL, Spark SQL, and U-SQL Azure Data Lake Analytics.
- Executed data ingestion to one or more Azure Services, including Azure Data Lake, Azure Storage, Azure SQL, and Azure DW. Processed the data efficiently within Azure Databricks.
- Created Pipelines in **Azure Data Factory** to Extract, Transform and load data from different sources like **Azure SQL, Blob storage, Azure SQL Data warehouse**, write-back tool and backwards.
- Developed JSON Scripts for deploying the Pipeline in **Azure Data Factory** (ADF) that process the data using the Sql Activity.
- Gained experience on developing SQL Scripts for automation purpose and Developed ETL Process using SPARK.
- Implemented of Data ingestion, Airflow Operators for Data Orchestration, and other related python libraries.
- Analyzed the SQL scripts and designed solutions to implement using PySpark.
- Utilized **Data bricks** notebooks for interactive analytics using Spark APIs
- Involved in building an Enterprise **Data Lake** using Data Factory and Blob storage, enabling other teams to work withmore complex scenarios and ML solutions.
- Used **Azure Data Factory,** SQL API and Mongo API and integrated data from MongoDB, MS SQL, and cloud (Blob, Azure SQL DB)
- Facilitated data for interactive **Power BI** dashboards and reporting purposes.
- Worked on Continuous Integration and Continuous Deployment (CI/CD) of the Applications into Azure Cloud.
- Executed the creation of various jobs in Jenkins, including Maven, Free-style, External, Pipeline, and Multi-configuration jobs.

Azure Data Engineer **Sep 2019 to Dec 2021**
Optum, India

- Led the development of batch processing applications that demanded functional pipelining utilizing **Spark** APIs.
- Involved in building a data pipeline and performed analytics using **AWS** stack (EMR, EC2, S3, RDS, Lambda, Glue, Redshift).
- Collaborated with client team to transform data and integrate algorithms and models into automated processes.
- Utilized **Spark's** in memory capabilities to handle large datasets on **S3 Data Lake**. Loaded data into **S3** buckets, then filtered and loaded into **Hive** external tables.
- Created and modified **SQL** stored procedures, functions, views, indexes, and triggers.
- Performed ETL operations using **Python, Spark SQL, S3** and **Redshift** on terabytes of data to obtain customer insights.
- Used programming skills in Python to build robust data pipelines and dynamic systems.
- Demonstrated a strong understanding of various AWS services, including **S3, EC2 IAM, RDS** and executed orchestrations and data pipelines utilizing AWS Step Functions, Data Pipeline, and Glue.
- Integrated data from a variety of sources, assuring that they adhere to data quality and accessibility standards.
- Engineered ETL (Extract/Transform/Load) processes, designing robust database systems, and crafting tools for both real-time and offline analytic processing.

- Leveraged Azure Cloud resources – Azure Data Lake Storage Gen2, Azure Data Factory, and Azures Data warehouse to build and operate a centralized cross-functional Data analytics platform
- Used knowledge in Hadoop architecture, HDFS commands and experience designing & optimizing queries to build data pipelines.

Data Engineer **Jan 2018 to Aug 2019**
Wipro, India

- Utilized **Spark's** in memory capabilities to handle large datasets on **S3 Data Lake**. Loaded data into **S3** buckets, then filtered and loaded into **Hive** external tables.
- Led the creation and modification of SQL stored procedures, functions, views, indexes, and triggers, leveraging strong hands-on experience.
- Performed ETL operations using **Python, SparkSQL, S3** and **Redshift** on terabytes of data to obtain customer insights.
- Involved heavily in setting up the CI/CD pipeline using **Jenkins**, **Terraform** and **AWS**
- Performed end-to-end Architecture & implementation assessment of various AWS services like Amazon EMR, Redshift, S3.
- Used AWS EMR to transform and move large amounts of data into and out of other AWS data stores and databases, such as Amazon Simple Storage Service (Amazon S3) and Amazon Dynamo DB
- Mastered various AWS services including S3, EC2 IAM, and RDS. Executed orchestration and data pipeline tasks utilizing AWS Step Functions, Data Pipeline, and Glue.
- Transformed the data using AWS Glue dynamic frames with PySpark, cataloged the transformed the data using Crawlers and scheduled the job and crawler using workflow feature.
- Designed and managed public/private cloud infrastructures using Confidential Web Services (AWS) which include **EC2, S3, Cloud Front, Elastic File System,** and IAM which allowed automated operations.
- Deployed Cloud Front to deliver content further allowing reduction of load on the servers.
- Created IAM policies for delegated administration within AWS and Configure **IAM Users / Roles /** Used AWS EMR to transform and move large amounts of data into and out of other AWS data stores and databases, such as Amazon Simple Storage Service (Amazon S3) and Amazon Dynamo DB.
- Developed Azure data factory Pipelines to pull data from Blob Storage account to **Azure SQL** Data warehouse, Azure SQL Database and Azure Data Lake storage.
- Implemented complete CI/CD process in new and existing ADF environment while making credentials secure in Azure Key Vault, and deployed production ready ADF resources using Azure pipeline releases.

Education

Master's in Business Analytics and Data Science **Jan 2022 to Dec 2023**
Wichita State University, 1845 Fairmount St, KS 67260

Projects

Data Integration Using Azure Services

- Utilized Twitter API to extract data from Twitter like tweets, user profiles, and hashtags and retrieve data in JSON format.
- Implemented data transformation tasks within Azure Data Factory (ADF) to process the raw Twitter data, using ADF's data flow capabilities to cleanse, filter, and enrich the extracted tweets, extracting essential information like tweet text, user location, timestamps, and user mentions.