Venkata Sai Nadella

Data Engineer

Phone No: (647)872-8327 **Email ID: venkatasainadella84@gmail.com**

Career Objective

Results-driven Data Engineer with 5+ years of experience in building scalable data pipelines, optimizing data architectures, and enabling analytics solutions. Adept in cloud platforms (Azure, AWS, GCP), SQL, Python, and big data tools like Spark and Kafka. Looking to contribute my expertise to a forward-thinking organization that values data-driven decision-making.

Profile Summary

- **5+** years of programming experience involved in all phases of Software Development Life Cycle (**SDLC**). Over 4+ Years of Big Data experience in building highly scalable data analytics applications.
- Expertise in building batch and real-time data processing systems leveraging **AWS** services (**S3**, **Redshift**, **EMR**, **Lambda**, **DynamoDB**) and **Azure** services (**ADLS, ADF**, **Databricks**, **Synapse Analytics**).
- Good Hand-on Experience in **ETL** processing, Migration and data processing using **AWS** services such as **EC2**, **Athena**, **Glue**, **Lambda**, **S3**, Relational Database Service (**RDS**) and other data-based services of **AWS**.
- Worked with **csv, Avro, Parquet** data to load into **Data frames** and do the analysis.
- Experience using various **Hadoop Distributions** (**Cloudera**, **MapR**, **Hortonworks**, **Azure**) to fully implement and leverage new **Hadoop** features. Used **Kafka** to load real-time data from multiple data sources into **HDFS**.
- Hands-on experience with **Spark**, **Databricks**, and **Delta Lake**.
- Experienced with **JSON** based **RESTful** web services, and **XML/QML** based **SOAP** web services and also worked on various applications using **python** integrated **IDEs** like **Sublime Text** and **PyCharm**.
- Strong knowledge on Architecture of Distributed systems and Parallel processing, In-depth understanding of MapReduce programming paradigm and **Spark** execution framework. Experience in Performance Monitoring, Security, Trouble shooting, Backup, Disaster recovery, Maintenance and Support of **Linux** systems.
- Involved in all phases of Software Development Life Cycle (**SDLC**) in large scale enterprise software using Object Oriented **Analysis** and Design. Experience in implementing **Azure** data solutions, provisioning storage account, **Azure Data Factory**, **SQL Server**, **SQL Databases**, **SQL Data warehouse**, **Azure Data Bricks** and **Azure Cosmos DB.**
- Clusters in **Databricks**, Managing the **Machine Learning** Lifecycle. Proficient in building **CI/CD pipelines** in **Jenkins** using pipeline syntax and groovy libraries.
- Good experience in automating the scheduled jobs in **Control-M** and **Airflow**.
- Have Extensive Experience in IT data analytics projects, Hands-on experience in migrating on premise **ETs** to **Google Cloud Platform** (**GCP**) using cloud native tools such as **BIG query, Cloud Data Proc, Google Cloud Storage, Composer.**
- Involved in using **Jenkins**, **Selenium**, **Docker** and **Kubernetes** clusters and power Shell scripting for efficient data management and decentralized access.
- Developed web-based applications using **Python**, **DJANGO, QT, C++, XML, CSS3, HTML5, DHTML**, **JavaScript** and **jQuery**. Data architectural expertise, including knowledge of **Data intake, Pipeline design**, **Hadoop** information architecture, Data modeling, Data mining, Sophisticated data processing, and **ETL** workflow optimization.
- Hands-on experience using **Snowflake** utilities, **SnowSQL**, **SnowPipe**, Big Data model techniques using **Python** / **Java**.
- Experience developing sophisticated **SQL** queries, procedures, and triggers using **RDBMS** like **Oracle** and **MySQL**.
- Experience in **Cisco Cloud Center** to more securely deploy and manage applications in multiple data center, private cloud, and public cloud environments. Worked on production support looking into logs, hot fixes and used Splunk for log monitoring along with **AWS** CloudWatch.

Education Details

Seshadri Rao Gudlavalleru Engineering College, India

- Bachelors of Technology, Computer Science and Engineering (Aug 2018 – April 2022)

Key Skills

AWS Services	S3, EC2, EMR, Redshift, RDS, Lambda, Kinesis, SNS, SQS, AMI, IAM, Cloud formation
Hadoop Components / Big Data	HDFS, Hue, MapReduce, PIG, Hive, HCatalog, HBase, Sqoop, Impala, Zookeeper, Flume, Kafka, Yarn, Cloudera Manager, Kerberos, PySpark Airflow, Kafka, Snowflake Spark Components
Databases	Oracle, Microsoft SQL Server, MySQL, DB2, Teradata
Programming Languages	Java, Scala, Impala, Python.
Web Servers	Apache Tomcat, WebLogic
IDE	Eclipse, Dreamweaver
NoSQL Databases	NoSQL Database (HBase, Cassandra, Mongo DB)
Methodologies	Agile (Scrum), Waterfall, UML, Design Patterns, SDLC
Currently Exploring	Apache Flink, Drill, Tachyon
Cloud Services	AWS, Azure, Azure Data Factory / ETL/ELT/SSIS Azure Data Lake Storage Azure Data bricks, GCP
ETL Tools	Talend Open Studio & Talend Enterprise Platform
Reporting and ETL Tools	Tableau, Power BI, AWS GLUE, SSIS, SSRS, Informatica, Data Stage

Employment History

Azure Data Engineer - Sun Life Financial Inc.

Toronto, Ontario, Canada **Jul 2024 - Current**

Sun Life is a financial services company. Design, build, and maintain data pipelines using tools like Azure Data Factory, Databricks, or Synapse Pipelines to ingest, transform, and load data from various source systems. Develop and optimize ETL/ELT processes for data warehousing and analytics purposes.

Responsibilities and Achievements:

- Create custom logging framework for **ELT** pipeline logging using Append variables in Data factory.
- Handled importing of data from various data sources, performed transformations using **B,** loaded data into **HDFS** and Extracted the data from **SQL** into **HDFS** using **Sqoop**.
- Involved in data validations and reports using **Power BI.** Implemented **Apache Sqoop** for efficiently transferring bulk data between **Apache** Hadoop and relational databases (**Oracle**) for product level forecast.
- Used **Kafka** functionalities like distribution, partition, replicated commit log service for messaging systems by maintaining feeds. Involved in loading data from rest endpoints to **Kafka**. Built robust data ingestion pipelines using **Logstash**, **Filebeat**, and **Kafka** Connect to stream real-time logs and events into **Elasticsearch** clusters.
- Performed data processing in **Azure Databricks** after data ingestion into **Azure** services such as **Azure** Data **Lake**, **Azure** Storage, **Azure SQL** DB, and **Azure SQL DW**. Created Data tables utilizing **PyQt** to display customer and policy information and add, delete, update customer records. Used **Python** library **Beautiful Soup** for web scrapping.
- Spark **SQL** to enable automated transformation of **RDD** case classes to schema **RDD** for both **Scala** and **Python** interfaces.
- Involved in various phases of Software Development Lifecycle (**SDLC**) of the application, like gathering requirements, design, development, deployment, and analysis of the application. Design and configure database, Back-end applications and programs. Managed large datasets using **Pandas** data frames and **SQL**.
- Used **Azure Data Factory** to ingest data from log files and business custom applications, processed data on Data bricks per day-to-day requirements, and loaded them to **Azure** Data Lakes.
- Implemented a continuous delivery (**CI/CD) pipeline** with **Docker** for custom application images in the cloud using **Jenkins**. Developed and automated data migration pipelines using **Python**, **Apache Airflow**, and **GCP** services, ensuring data consistency and minimizing downtime during cutover.
- Responsible for implementing monitoring solutions in **Ansible, Terraform, Docker**, and **Jenkins**.
- Implemented Synapse Integration with **Azure Databricks** notebooks which reduce about half of development work.
- Conducted Performance tuning and optimization of **Snowflake** data warehouse, resulting in improved query execution times and reduced operational costs. Used **Python, R, SQL** to create Statistical algorithms involving Multivariate Regression, Linea Regression, Logistic Worked on **Kafka** publishing the messages for further downstream systems.

Technologies Used: Airflow, Apache, Azure, Azure Data Lake, CI/CD, Data Factory, Docker, EC2, Elasticsearch, ELT, Factory, GCP, HDFS, Jenkins, Kafka, Lake, Oracle, Pandas, Power BI, Python, Scala, Snowflake, Spark, Spark SQL, SQL, Sqoop

AWS Data Engineer - Astellas Pharma Canada

Markham, Ontario, Canada **Jun 2023 - Jun 2024**

Astellas Pharma Canada, Inc., is the Canadian affiliate of Tokyo-based multinational pharmaceutical company. I designed, developed, and implemented the data integration and ETL/ELT processes using AWS Glue, AWS Data Pipeline, and Step Functions. Optimized the data pipelines for performance, efficiency and data quality.

Responsibilities and Achievements:

- Provisioned high availability of **AWS EC2** instances, migrated legacy systems to **AWS**, and developed Terraform Plugins, modules, and templates for automating **AWS** infrastructure.
- Designed and implemented **ETL** (Extract, Transform, Load) processes using **C# to** cleanse, transform, and enrich raw data, ensuring its quality and compatibility with downstream analytics and reporting systems.
- Developed serverless data processing functions using **AWS Lambda**, triggered by **S3** events, **DynamoDB** streams, and **API Gateway**. Led migration to **AWS**, leveraging **Amazon Redshift** for data warehousing and utilizing **HiveQL** for reporting, reducing data retrieval and processing time by 30%.
- Actively involved in designing and developing data ingestion, aggregation, and integration in the Hadoop environment.
- Have worked on partition of **Kafka** messages and setting up the replication factors in **Kafka Cluster**.
- Converted and parsed data formats using **PySpark Data Frames**, reducing time spent on data conversion and parsing by 40%. Involved in various phases of Software Development Lifecycle (**SDLC**) of the application, like gathering requirements, design, development, deployment, and analysis of the application.
- Used **Django** evolution and manual **SQL** modifications were able to modify **Django** models while retaining all data, while site was in production mode. **Python machine learning** techniques were used to anticipate user order amounts for certain goods, with automated recommendations provided via **Kinesis Firehose** and **S3 data lake**.
- Worked on **CI/CD** tools like **Jenkins**, **Docker** in **Devops Team** for setting up application process from end-to-end using Deployment for lower environments and Delivery for higher environments by using approvals in between.
- Proficient in utilizing **Sqoop** for the seamless transfer of data from diverse relational databases to **Hadoop** Distributed File System. Designed and executed data migration strategies for relational databases (**MySQL, PostgreSQL**) to cloud-native solutions like **BigQuery** and Cloud **SQL**, reducing query latency by 40%.
- Building/Maintaining **Docker** container clusters managed by **Kubernetes Linux**, **Bash**, **Git**, **Docker**.
- Optimized **Elasticsearch** cluster performance through shard tuning, heap memory management, refresh interval adjustments, and query profiling. Provisioned high availability of **AWS EC2** instances, migrated legacy systems to **AWS**, and developed Terraform plugins, modules, and templates for automating **AWS** infrastructure.
- Conducted query optimization and performance tuning tasks, such as query profiling, indexing, and utilizing Snowflake's automatic clustering to improve query response times and reduce costs.
- Developed remote integration with third-party platforms by using **RESTful web services.**

Technologies Used: API, AWS, BigQuery, CI/CD, Cluster, Docker, DynamoDB, EC2, Elasticsearch, ETL, Gateway, Git, Hive, HiveQL, Jenkins, Kafka, Kubernetes, lake, Lambda, Linux, MySQL, PostgreSQL, PySpark, Python, Redshift, S3, Snowflake, Spark

GCP Data Engineer - Aditya Birla Sun Life Asset Management

Mumbai, India **Jan 2022 - Mar 2023**

Aditya Birla Sun Life Asset Management Company Ltd. is an investment managing company. Managed the data lifecycle in GCP Cloud Storage, including ingestion, archival, and security policies. Integrated Cloud Storage with on-prem or other cloud sources using Dataflow or custom ingestion tools.

Responsibilities and Achievements:

- Have written hive and spark queries using optimized ways like using window functions, customizing Hadoop shuffle & sort parameter. Utilized **Google Cloud Shell** for rapid development and testing of data engineering workflows, enabling instant access to **GCP** resources without local setup.
- Ensured data quality and report accuracy by implementing validation scripts and schema checks in the pipeline feeding Data Studio. Optimized **Spark** jobs on **Dataproc** clusters by tuning memory and executor settings, improving job performance by 30% and lowering cluster cost by 20%.

- Achieved 70% faster **EMR** cluster launch and configuration, optimized **Hadoop** job processing by 60%, improved system stability, and utilized Boto3 for seamless file writing to **S3 bucket**. Developed and optimized **Spark** jobs on **Databricks** clusters to process large-scale datasets (TBs+), improving runtime by 30%.
- Used **Power BI** as a front-end **BI tool** to design and develop dashboards, workbooks, and complex aggregate calculations.
- Developed User-Defined Functions (**UDFs**) in **Scala** and **Pyspark** to meet specific business requirements.
- Used R and **Python** for Exploratory **Data Analysis** to compare and identify the effectiveness of the data.
- Experienced in **Google Cloud components**, **Google container** builders and **GCP** client libraries and **Cloud SDK'S.**
- Pipelines were created in **Azure Data Factory** utilizing Linked **Services** to extract, transform, and load data from many sources such as **Azure SQL** Data warehouse, write-back tool, and backwards. Ingested raw data from Cloud Storage, Pub/Sub, and third-party **APIs** into **BigQuery** for downstream analytics and machine learning workflows.
- Used DataStax **Spark** connector which is used to store the data into **Cassandra** database or get the data from **Cassandra** database. Have used **T-SQL** for MS **SQL** server and **ANSI SQL** extensively on disparate databases.
- Used **Sqoop** import/export to ingest raw data into **Google Cloud Storage** by spinning up **Cloud Dataproc cluster**.
- Integrated Dataflow with **BigQuery**, **Cloud Storage**, and **Firestore** to support analytics, machine learning, and operational use cases. Involved in setting up of **Apache Airflow service** in **GCP**.

Technologies Used: Airflow, Analysis, Apache, API, Azure, BigQuery, Cassandra, Data Factory, EMR, Factory, GCP, Power BI, Python, S3, Scala, SDK, Services, Spark, SQL, Sqoop

Data Engineer - Blue Dart

Mumbai, India **Jul 2020 - Dec 2021**

Blue Dart Express is an Indian logistics company that provides courier delivery services. I utilized the Spark on Cloud Dataproc, Hadoop, and Apache Beam for big data processing and advanced analytics. Ensured the performance optimization of dashboards consuming BigQuery datasets.

Responsibilities and Achievements:

- Performed **ETL** to move the data from source s oystem to destination systems and worked on the Data warehouse. Involved in database migration methodologies and integration conversion solutions to convert legacy **ETL** processes into **Azure** Synapse compatible architecture.
- Developed **T-SQL** scripts to create, modify, and manage **Azure SQL Database** objects such as tables, indexes, and views.
- Employed **Hadoop scripts** to manipulate and load data from the **Hadoop File System.**
- Experience in using **Kafka** as a messaging system to implement real-time Streaming solutions using **Spark** Streaming
- Developed **Databricks ETL** pipelines using notebooks, **Spark** Data frames, **Spark SQL** and **Python** scripting.
- Implemented data transformations and enrichment using **Apache Spark Streaming** to clean and structure the data for analysis. Responsible for building scalable distributed data solutions using **Hadoop**.
- Actively Participated in all phases of the Software Development Life Cycle (**SDLC**) from implementation to deployment.
- Developing **Spark** scripts, UDFS using both **Spark** DSL and **Spark SQL** query for data aggregation, querying, and writing data back into **RDBMS** through **Sqoop**. Utilized **Elasticsearch** and **Kibana** for indexing and visualizing the real-time analytics results, enabling stakeholders to gain actionable insights quickly.
- Used **Azure** Key vault as central repository for maintaining secrets and referenced the secrets in **Azure Data Factory** and also in **Databricks** notebooks. Worked on scheduling all jobs using **Airflow** scripts using **Python**. Adding different tasks to **DAG's** and dependencies between the tasks.
- Implemented automated Data pipelines for Data migration, ensuring a smooth and reliable transition to the Cloud environment. Skilled in monitoring servers using **Nagios, Cloud watch** and using **ELK Stack- Elastic search** and **Kibana.**
- Experience in creating **Kubernetes** replication controllers, Clusters and label services to deployed Microservices in **Docker**.
- Develop metrics based on **SAS** scripts on legacy system, migrating metrics to **Snowflake** (Azure).
- Used **Redshift** Spectrum with wide range of data formats, like **Parquet, ORC, CSV, JSON**, etc

Technologies Used: Azure, Azure SQL Database, CI/CD, Cluster, Data Factory, Docker, Elasticsearch, ETL, Factory, Git, Jenkins, JS, Kafka, Kubernetes, Python, RDBMS, Redshift, SAS, Snowflake, Spark, Spark SQL, Spark Streaming, SQL, Sqoop