Manish Badugu

Email: badugu.manishh@gmail.com

Phone: (901)-286-5497

PROFESSIONAL SUMMARY:

- Data Engineer professional with around 11 years of IT experience in complete life cycle of software development using Object Oriented analysis and design using Cloud tools, Big Data Technologies, Spark.
- Experienced in SQL, Python, Azure, AWS, Snowflake, Google Cloud Platform, Tableau and MS Tools.
- Good experience with Big Data technologies including Spark, HDFS, Yarn, Pig, Hive, Sqoop, Flume and Kafka.
- Core group member and responsible for transitioning a large sized project from Waterfall to Agile.
- Strong experience in Extraction, Transformation and Loading (ETL) data from various sources into Data Warehouses and Data Marts using Informatica Power Canter, Power Exchange, Power Connect as ETL tool on Oracle, DB2 and Teradata Databases.
- Expertise in creating clusters in Google Cloud and manage the clusters using Kubernetes (k8s). Using Jenkins to deploy code to Google Cloud, create new namespaces, creating Docker images and pushing them to container registry of Google Cloud.
- Experience building data pipelines on GCP using Dataflow, Pub/Sub, and BigQuery for scalable, event-driven architectures.
- Applied machine learning pipelines (feature engineering, data preparation) to support predictive analytics and AI-driven use cases.
- Integrated ERP systems (SAP/Oracle/Finance platforms) with modern data platforms for enterprise reporting and analytics.
- Strong experience working with Snowflake advanced features such as clustering, time travel, zero-copy cloning, and performance optimization.
- Built and maintained Hadoop ecosystem workflows using HDFS, Hive, and MapReduce for large-scale batch processing.
- Experience designing data lakehouse architectures combining Snowflake/BigQuery with object storage.
- Exposure to MLOps concepts, including data versioning, pipeline reproducibility, and model data validation.
- Implemented cross-cloud data integration strategies (GCP ↔ AWS ↔ Azure) for unified data platforms.
- Experience enabling self-service analytics by structuring data for BI and ML consumption layers.
- Strong understanding of data lifecycle management, including archival, retention policies, and cost optimization strategies.
- Experienced in migrating applications to AWS and application deployment in the cloud (AWS) with CI/CD tools like Jenkins.
- Experience working with varied forms of data infrastructure inclusive of relational databases such as SQL, Hadoop, Spark, and column-oriented databases such as MySQL.
- Experienced with Python programming and Python libraries for data science including NumPy, Pandas, and SciPy etc.
- Expert in writing SQL queries and managing MS access and MS SQL server database for data entry, storage and generation of reports to facilitate management decision making processes.
- Extract, transform and load data using Access Database and conduct data analysis using Process Modelling, Data Mining, Data Modelling and Data Mapping. Develop and maintain blueprints for Access database bill readers.
- Experience in collaborating with developers as a group lead in designing Access ETL, data models and database architecture using data warehousing concepts.
- Expertise in designing SSIS Packages to extract, transfer, load (ETL) existing data into SQL Server from different environments for the SSAS cubes.
- Demonstrated experience in building and maintaining reliable and scalable ETL on big data platforms.
- Proficiency in data warehousing inclusive of dimensional modelling concepts and in scripting languages like Python, Scala, and JavaScript.
- Substantial experience working with big data infrastructure tools such as Python and Redshift also proficient in Scala, Spark, and Spark Streaming.
- Experience working on implementing CRUD operations using NoSQL Rest APIs.
- Strong interpersonal communication skills and ability to work independently as well as in a group setting.
- Highly organized with the ability to work collaboratively with all the team members to ensure high-quality products and manage multiple projects.
- Strong experience in Extraction, Transformation and Loading (ETL) data from various sources into Data Warehouses and Data Marts using Informatica Power Canter, Power Exchange, Power Connect as ETL tool on Oracle, DB2 and Teradata Databases.
- Extensive experience in Amazon Web Services like S3, CloudFront, EC2, AWS Backup, RDS, Elastic Load Balancing, SQS, SNS, AWS IAM, AWS Cloud Watch and Redshift.
- Experience integrating enterprise ERP systems (SAP/Oracle) with cloud data platforms to enable centralized, analytics-ready datasets for business reporting and financial insights.
- Hands-on exposure to AI/ML data pipelines, including feature engineering, data preprocessing, and preparing scalable datasets for model training and predictive analytics.

TECHNICAL SKILLS:

Programming Languages	SQL, Python, C, Java, HTML, Scala, JavaScript, Shell Scripting.
Python Libraries	NumPy, Pandas, NLTK, Matplotlib, Seaborn, Scikit-learn, TensorFlow, PyArrow.
Big Data Tools	Apache Spark, Apache Hadoop, Apache Kafka, Hive, Spark SQL, MapReduce, Apache Flink, Delta Lake / Iceberg.
Microsoft Tools	Microsoft Project, Access, Excel, Word, Visio, PowerPoint, SharePoint, Publisher, Outlook, MS office Suite, Power Platform, Excel Power Query.
Applications and Tools	JIRA, Power BI, Tableau, Data Warehousing, AWS, GitHub, ERP System, Google Cloud Platform, Snowflake, Apache Airflow, Databricks, Docker, Terraform, ERP Integration.
IDEs	Eclipse, IntelliJ, Atom, Visual Studio.
Databases	MS SQL Server, Oracle, MySQL Workbench, Amazon Redshift, Teradata, PostgreSQL, Snowflake, Google BigQuery.
Operating Systems	Windows, UNIX, LINUX.
Analysis/Methodologies	Agile/Scrum, Waterfall, Kanban, RUP, Cost/Benefit Analysis, Test Drive Development, Data Modeling, Data Governance & Data Quality Frameworks, MLOps Concepts, Event-Driven Architecture, CI/CD.

PROFESSIONAL EXPERIENCE:

American Airlines, Dallas, TX **Jun 2024 - Present**

Data Engineer

Responsibilities:

- Extract, transform and load data using Access Database and conduct data analysis using Process modelling, Data Mining, Data Modelling and Data Mapping. Develop and maintain blueprints for Access database bill readers.
- Develop highly optimized Spark applications to perform data cleansing, validation, transformation, and summarization activities.
- Built streaming pipelines on GCP using Pub/Sub and Dataflow for real-time ingestion and transformation of high-velocity data.
- Designed feature engineering pipelines to support machine learning models, including data normalization, aggregation, and enrichment.
- Implemented Snowflake performance tuning techniques (clustering keys, query pruning, warehouse sizing) to optimize query execution.
- Integrated enterprise ERP data sources (SAP/Oracle) into cloud data platforms for unified reporting and financial analytics.
- Designed data lakehouse architecture combining structured and semi-structured data using Snowflake and cloud storage.
- Built reusable ingestion frameworks to handle ERP and transactional system data with schema evolution support.
- Implemented data validation layers to ensure ML-ready datasets met consistency and accuracy requirements.
- Developed cross-region data pipelines on GCP to improve data availability and disaster recovery readiness.
- Enabled downstream AI use cases by creating curated datasets for training and inference workflows.
- Applied cost optimization strategies in Snowflake and GCP by managing compute usage and storage lifecycle.
- Work on NoSQL databases like HBase and imported data from MySQL and processed data using Hadoop Tools and exported to Cassandra NoSQL database.
- Develop and design Security Framework to provide fine grained access to objects in AWS S# using AWS Lambda, DynamoDB.
- Build and architect multiple Data pipelines, end to end ETL and ELT process for Data Ingestion and transformation in AWS and Spark.
- Create multiple scripts to automate ETL/ELT process using PySpark from multiple sources
- Develop Airflow Bigdata jobs to load large volume of data into S3 data lake and then into Snowflake.
- Perform end-to-end architecture and implement assessment of various AWS services like Amazon EMR, Redshift, and S3.
- Work in different Python data science libraries like NumPy and Pandas and did a POC on Sentiment Analysis.
- Involve in converting legacy system queries into Spark and SQL transformations using Spark RDDs, Python and SQL.
- Work on Tableau to create complex Visualization and support multiple dashboards based on business requirement.
- Work with AWS stack S3, EC2, EMR, Athena, Glue, Redshift, DynamoDB, IAM, and Lambda.
- Responsible for design and development of Spark SQL Scripts based on Functional Specifications.
- Migrating Objects from Teradata, SQL Server to S3 and then on Snowflake.

- Involved in designing optimizing Spark SQL queries, Data frames, import data from Data sources, perform transformations, perform read/write operations, save the results to output directory into HDFS/AWS S3.
- Develop Airflow DAG's for scheduling the workflows/views responsible for creating tables and updating the existing data.
- Create Partitioned Hive tables and worked on them using HiveQL.
- Loading Data into HBase using Bulk Load and Non-bulk load.
- Building data pipeline ETLs for data movement to S3, then to Redshift.
- Involve in continuous Integration of applications using Jenkins.
- Implement reporting Data Warehouse with online transaction system data.
- Intake happens through Sqoop and Ingestion happens through Map Reduce, HBASE.
- Working on migrating Data pipelines to in-house MAP-AIRFLOW from legacy Airflow cluster.
- Worked in snowflake to create and Maintain Tables and views.
- Strong experience in writing scripts using Python API, PySpark API, and Spark API for analyzing the data.
- Communicate with peers and supervisors routinely, document work, meetings, and decisions.

Environment: Python, Spark, Dataflow, AWS, Hadoop, HDFS, Sqoop, MYSQL, Oracle, ETL Methods, Linux, Data Studio, Airflow, Concord, Scala, Dataproc, Compute Engine, Pub/Sub, Dataflow, Big Query

Verizon, NYC NY **Jan 2023 - May 2024**

Data Engineer

Responsibilities:

- Designed and built a very complex pipelines which can bring data from various data sources and land it in ADLS location.
- Designed, developed, tested, implemented and supported Data Warehousing ETL using Abinitio Technology.
- Dealt with both ETL and ELT architectures using Data Factory, Databricks and mapping dataflow.
- Processed and loaded bound and unbound Data from Google pub/sub topic to BigQuery using Cloud Dataflow with Python.
- Developed Hadoop-based batch processing pipelines using HDFS, Hive, and MapReduce for large-scale historical data processing.
- Built hybrid architectures integrating Hadoop ecosystems with modern cloud warehouses like Snowflake and BigQuery.
- Designed pipelines to extract and transform ERP system data into analytical data models for business intelligence.
- Implemented ML data pipelines to support model training workflows, including dataset versioning and preprocessing.
- Leveraged BigQuery advanced capabilities such as partitioning and clustering for high-performance analytics.
- Created data contracts and schema governance strategies for consistent integration across ERP and external systems.
- Designed incremental data loading strategies for large enterprise datasets to reduce processing overhead.
- Built data quality frameworks specifically tailored for ML pipelines to ensure feature reliability.
- Integrated AI/ML workflows with data pipelines, enabling automated data preparation for predictive models.
- Migrated legacy Hadoop workloads to cloud-native platforms (GCP/Snowflake) while maintaining performance and reliability.
- Worked real time streaming with Kafka as a data pipeline using spark streaming module.
- Controlled and granted database access and migrated on premise databases to Azure Data Lake store using Azure Data Factory.
- Developed Python scripts to automate the ETL process using Apache Airflow and CRON scripts in the UNIX operating system as we ll.
- Integrated Talend Open Studio with Hive, Spark and MySQL.
- Created custom Denodo views by joining tables from multiple data sources.
- Wrote Scala program for Spark transformation in Dataproc.
- Heavily involved in testing Snowflake to understand best possible way to use the cloud resources.
- Created the hive external tables for developer reference on top of ADLS locations.
- Developed Spark jobs to parse the JSON, CSV, Avro and Parquet data.
- Achieved performance improvement of existing jobs by fine tuning Spark Jobs and improved spark application 3X times.
- Created multiple VPC's and public/private subnets, Route tables, Route Tables Security groups and Elastic Load 3X times.
- Extract Transform and Load 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. Data Ingestion to one or more Azure Services - (Azure Data Lake, Azure Storage, Azure SQL, Azure DW) and processing the data in Azure Databricks.
- Handled real time streams data using EventHub's with structured Streaming.
- Implemented Spark using Scala and Spark SQL for faster testing and processing of data.
- Designed and developed ETL Processes in AWS Glue to migrate Campaign data from external sources like S3, ORC/Parquet/Text Files into AWS Redshift.

- Scheduled different Snowflake jobs using NiFi.
- Created firewall rules to access Google Data proc from other machines.

Environment: ETL, Azure, Apache Beam, Snowflake, Cloud Dataflow, Cloud Shell, Cloud SQL, MySQL Terraform, Snowflake, Postgres, Python, Scala, Spark, Hive, Spark-SQL

DaVita, El Segundo, CA **Nov 2019 - Dec 2022**

Data Engineer

Responsibilities:

- Built and architected multiple Data pipelines, end to end ETL and ELT process for Data ingestion and transformation in GCP and coordinate task among the team.
- Coordinated with the Team Manager & the team for various complex business requirements through SCRUM Agile process.
- Used Rest API with Python to ingest Data from and some other site to BigQuery.
- Implemented and maintained ETL pipeline for processing raw sales and campaign data using Java 8, Shell Scripting, Hive, Spark and power of Distributed Computing.
- Worked on importing and exporting data from Snowflake, Oracle and DB2 into HDFS and HIVE using Sqoop for analysis, visualization and to generate reports.
- Set-up databases in AWS using RDS, storage using S3 bucket and configuring instance backups to S3 bucket.
- Built Scala and Spark based configurable framework to connect common Data sources like MYSQL, Oracle, Postgres, SQL Server, Salesforce, Big query and load it in Big query.
- Responsible for optimization and troubleshooting, test case integration into CI/CD pipeline using Docker images
- Built data pipeline ETLs for data movement to S3, then to Redshift.
- Created IAM policies for delegated administration within AWS and Configure IAM Users / Roles / Policies to grant fine - grained access to AWS resources to users.
- Designed and implemented ETL pipelines between from various Relational Databases to the Data Warehouse using Apache Airflow.
- Tested the ingested data to maintain consistency and parity with the source data.
- Created and maintained fully automated CI/CD pipelines for code deployment.
- Involved in design, development, testing, and implementation of the process systems, working on iterative life cycles business requirements, and creating Detail Design Document.
- Participated in scrum meetings and co-ordinate the development process with various other project modules.
- Used python operators such as bash operator, Hive operators and customized spark operators to build DAGS and schedule Spark Applications using Airflow.

Environment: ETL, Agile, Java 8, Shell Scripting, Hive, MYSQL, Oracle, Postgres, SQL Server, Scala, Spark, Python, AWS, Apache Airflow, Data Warehouse.

Bodh tree Consulting Ltd **Aug 2016 - Nov 2018**

Data Engineer

Responsibilities:

- Interpreted data, analyzed results using statistical techniques and provided ongoing reports.
- Developed and implemented databases, data collection systems, data analytics and other strategies that optimize statistical efficiency and quality.
- Involved in the Data profiling activities and come out with generic test cases.
- Carried deployments and builds on various environments using continuous integration tool Jenkins.
- Involved in designing of APIs for the networking and cloud services and Leveraged spark (PySpark) to manipulate unstructured data and apply text mining on user's table utilization data.
- Developed Spark/Scala, Python for regular expression (regex) project in the Hadoop/Hive environment with Linux/Windows for big data resources.
- Used Spark API over Hadoop YARN as execution engine for data analytics using Hive. Used Spark API over Hortonworks Hadoop YARN to perform analytics on data in Hive.
- Ingested data from various data sources into Hadoop HDFS/Hive Tables using SQOOP, Flume, Kafka.
- Integrated Tableau with Hadoop data source for building dashboard to provide various insights on sales of the organization.
- Created multi-node Hadoop and Spark clusters in AWS instances to generate terabytes of data and stored it in AWS HDFS.
- Designed Data Quality Framework to perform schema validation and data profiling on Spark (PySpark).
- Implemented advanced procedures like text analytics and processing using the in-memory computing capabilities like Apache Spark written in Scala.
- Used Pandas API to put the data as time series and tabular format for easy timestamp data manipulation and retrieval.
- Managed large datasets using Panda’s data frames and MySQL.

- Developed the required XML Schema documents and implemented the framework for parsing XML documents.
- Involved in importing the real time data to Hadoop using Kafka and implemented the Oozie job for daily imports.

Environment: Python, Spark, Pandas, NumPy, PySpark, Hadoop, Scala, HDFS, Hive, MySQL, Kafka, Linux/Windows, SQL, Impala, Spark.

Max Health Care, India **Apr 2014 - Jul 2016**

Software Engineer

Responsibilities:

- Gathered business requirements and converted them into new T-SQL stored procedures in visual studio for database project.
- Interpreted data, analyzed results using statistical techniques and provided ongoing reports.
- Prepared a claim data for risk adjustment by extracting, cleaning, and merging from different tables using SQL.
- Built and maintained SQL scripts, indexes, and complex queries for data analysis and extraction for varies projects.
- Wrote and executed SQL queries using Query Analyzer to provide custom reports to marketing and sales.
- Developed and implemented data collection systems and other strategies that optimize statistical efficiency and data quality.
- Contributed to the spring boot based multi-tier web application using Java 8, Spring Boot, Microservices • Performed and automated SQL Server version upgrades, patch installs and maintained relational databases.
- Modified and maintained SQL Server stored procedures, views, ad-hoc queries, and SSIS packages used in the search engine optimization process.
- Updated existing and created new reports using Microsoft SQL Server Reporting Services. • Created files, views, tables, and data sets to support Sales Operations and Analytics teams
- Monitored and tuned database resources and activities for SQL Server databases.
- Designed and developed the REST style web services using Python and Flask, Postgres Database.
- Wrote complex SQL queries and PL/SQL functions.
- Developed the Command Line Interface (CLI) tool for red hat Linux.
- Used Python's XML parser architectures (SAX) and DOM API for tracking small amounts of data without requiring the DB.
- Created the Linux Services to run REST web services using Shell script.
- Built the RPM Package for the Product with the upgrade features support.
- Designed and developed the test cases for REST API, Involved REST API test framework development.
- Used Python Library Beautiful Soup for web scraping
- Designed and developed the test cases for CLI automation using Python.
- Performed Unit testing and developed the unit test cases using PyUnit framework.
- Used Jenkins to deploy web services and run unit tests, REST API test.
- Involved in automation of creations of VLAN, Trunk port and Routing.
- Created branch and committed the code changes to Master branch using SVN version control and commands in Linux.

Environment: Python, Flask, BeautifulSoup, Java, PL/SQL, Linux, HTML, XHTML, CSS, AJAX, JavaScript, SSRS, SSIS, PyUnit, REST API, Shell Scripting, SQL, T-SQL, Query Analyzer, SVN, SQL Server.