VIJAY SANJAY NALE

Big Data Developer

Email : vsnale.04@gmail.com
Mob No: 8208078704
LinkedIn
Github

CAREER OBJECTIVE

Dedicated Big Data professional with **3.7 years** of hands-on experience in analyzing large datasets, implementing data-driven solutions, and optimizing data processes. Proficient in leveraging cutting-edge technologies and tools such as Hadoop, Spark, and Python to extract valuable insights and drive business growth. Seeking to utilize my expertise in data analysis, and data engineering to contribute to a forward-thinking organization's success in harnessing the power of data for strategic decision-making and innovation.

EDUCATION

Master of Computer Application (C.B.S.E Board) (July '16 - Apr '19)
Shivaji University, Kolhapur: **8.74 CGPA**

KEY SKILLS

Programming Languages	Python, Scala
Hadoop Platform	Apache Hadoop, Cloudera Distributed Environment
Hadoop Ecosystem/ Framework	Sqoop, Hive, Kafka, Airflow, HBase, Apache Spark (Spark Streaming, Spark SQL)
Database	SQL, PostgreSQL, Mysql
Tools	Pycharm, WinSCP, Putty, PyCharm, Jupyter, VS Code
Version controls	GitHub, Git
Methodologies	Agile, CICD, SDLC, Jira

PROFESSIONAL EXPERIENCE

A) Software Engineer, Data - Maveric Systems Limited, Pune (Oct '23 – Present)

- Designed and implemented **ETL processes** using Apache Spark and PySpark to efficiently extract and transform data from multiple sources, resulting in a **30% reduction** in data processing time.
- Developed and maintained **Hive queries** and scripts for data analysis and reporting, enabling business stakeholders to make data-driven decisions.
- **Optimized Spark jobs** by fine-tuning configurations and implementing **performance-enhancing techniques**, resulting in **a 20% improvement** in job execution time.
- Managed data storage and retrieval using **HDFS**, ensuring data integrity and high availability of critical datasets.
- Experience in developing spark application using **spark-SQL** in **databricks** for data extraction,
- transformation and aggregation from multiple file format for **analyzing & transforming** the data.
- Used **Spark-SQL** to process the data and to run on **Spark engine**.

B) Associate - Bajaj Finserv Ltd., Pune (Sep '21 – Present)

- Experienced with the tools in **Hadoop Ecosystem included Hive, HDFS, Map Reduce, Sqoop, Spark.**
- Excellent with the tools in Hadoop Ecosystem including **Resource Manager, Node Manager, Name Node,**
- **Data Node** and Map Reduce paradigm.
- Worked on **Hive Queries** for creating and querying Hive tables to retrieve useful **analytical information**.

- Used **Sqoop to import large data** from traditional **RDMS to HDFS** and also handled incremental Sqoop.
- Created **partitions, buckets** in **Hive**, Handled the **integration of Spark with Hive.**
- **Optimized the existing pipelines** which resulted in the reduction of execution time **by 10 times.**
- Experience of AWS component like **S3, EC2, EMR, AWS Athena, Crawler, Redshift, Glue.**
- Developed **Complex SQL queries** to retrieve data from the database as well as for performance tuning.
- Used **Airflow** to automate jobs.
- Experience in developing spark application using spark-SQL in databricks for data extraction, transformation and aggregation from multiple file format for analyzing & transforming the data.
- Used Spark-SQL to process the data and to run on Spark engine.

C) Data Analytics - Acceline Digital Media Pvt Ltd (Oct '19 – Sep '20)

- **Excellent SQL skills** for daily testing of data using various SQL queries.
- Reviewed analyze and implement necessary changes in appropriate areas to enhance and improve existing system, Perform DAX queries using Power BI tools.
- Designed, developed and tested various Power BI visualizations for dashboard and ad-hoc reporting
- solutions by connecting from different data sources and databases.

PROJECT

A. Data Engineering Pipeline for Bank Dataset Processing.

Spearheaded the development of a comprehensive data engineering pipeline to process and analyse a CSV dataset containing banking transactions. Leveraged a combination of Python, **Hive and Apache Spark** to perform **data cleaning, transformation**, and aggregation tasks, ensuring the integrity and quality of the data for downstream analysis.

- Reduced data processing time by **50% through parallelization** and optimization techniques in Spark.
- Designed and implemented a **scalable Hive data warehouse**, improving data accessibility and query performance.
- Developed interactive dashboards in Power BI, enabling stakeholders to gain insights and make data-driven decisions effectively.
- Technology Used : PySpark, Hive, SQL, Power BI

B. Analyze Insurance Data.

The theme of project is to **perform the analysis on huge dataset** which are generated by websites, app and other sources like employee portal etc. In this Application-Quotes level data get generated on daily basis. **Business look for number** so here we are **analyzing multiple things** like **how many payment success and their gross premium** are made by current month and also by each Category and from which UTM source.

- Daily **3 million data** records are processed, data is received multiple data producers.
- Importing data using **Sqoop into hive** and HDFS from existing SQL server (like Application-Quote Level, Order data, Customer data, Product data etc.).
- Write Script for processing data through **Spark Core and Spark SQL** to generate business report.
- Technology Involved: Python, PySpark, Hive, Sqoop, Airflow.

Certificate

- Machine learning Master Certification from iNeuron.ai
- Big Data Engineering from Trendytech.
- Spark and Python for Big Data with PySpark (Udemy), 2024
- Hive to Advanced Hive: Real Time usage (Udemy), 2024
- The Ultimate MySQL Bootcamp (Udemy), 2024