Role and Responsibilities:

• Lead data science projects from start to finish, including data collection, processing, analysis, visualization and prediction modelling.

• Analyze large and complex datasets using statistical and machine learning techniques.

• Develop and implement data-driven solutions to solve complex business problems.

• Design and develop Tableau Dashboards and reports to provide actionable insights to stakeholders.

• Use GCP to extract, transform, and load large and complex datasets.

• Collaborate with business stakeholders to identify opportunities for data-driven improvements.

• Mentor junior team members and provide technical guidance and support.

• Stay up-to-date with emerging trends and technologies in data analytics/science.

HADOOP GCP DATA GOVERNANCE LEAD DATA SCIENTIST

Candidate Profile:

• At least 5 years of experience as data scientist with expertise in Deep Learning, Tableau, Hadoop, GCP and SQL. Looking for folks who can drive adoption and not pick up skill on the job.

• Strong proficiency in programming language Python.

• Hands-on experience in automating the data science models and job scheduling.

• Hands-on for building the models and solid knowledge on Statistics.

• In-depth knowledge in Deep Learning, Time Series forecasting, Regression Techniques, Clustering and data filtering techniques.

• Ability to build relevant impactful visuals for the business stakeholders.

• Improving data governance and quality, increasing the reliability of data.

• Proven skills in translating analytics output to actionable recommendations.

• Strong verbal and written communication skills.

• Self-starter and independent worker.

• Capable of applying judgment to plan and execute your tasks.

Education:

Bachelor's degree or equivalent and work experience in Data Science, GCP, Python and SQL.

**Deep Learning, Time Series forecasting, Regression Techniques, Clustering  
Strong proficiency in programming language Python**

5 years of experience, multiple projects, scikit-learn, pytorch, tensor flow, NLP: NLTK, tokenizers

Comparative Analysis of ARIMA & Neural Network Models for forecasting International Flight Departure

Decision Trees, Random Forest and SVM

Integrated data into Power BI to create 10 dynamic dashboards and PowerPoint decks reducing report generation time by 20%.

**Lead data science projects from start to finish, including data collection, processing, analysis, visualization and prediction modelling.**

Data Governance, ETL

Analyze large and complex datasets using statistical and machine learning techniques.

Develop and implement data-driven solutions to solve complex business problems.

Customer segments, purchase behaviours, data filtering

**GCP -data proc, HADOOP**

**Team leader**

**Strong verbal and written communication skills Self-starter and independent worker.**

I am Prathamesh Joshi. I have more than 5 years of experience in Data Science field particularly in Data Science, Machine Learning and Reporting roles. I have recently completed my Masters in Data Science from University of Michigan Ann Arbor. Previously I have an experience of working at Axtria, India. I have 2 internship experiences in Data Analytics field.

I have completed Azure fundamental certifications and also LinkedIn learning on Healthcare industry.

Here's a comparison table between your **Olympic Data Analysis Project on Azure Synapse, Databricks, and PySpark** and a similar project using **GCP and Hadoop**:

| **Aspect** | **Azure Synapse, Databricks, and PySpark** | **GCP with Hadoop (Dataproc)** | **Similarities** |
| --- | --- | --- | --- |
| **Data Storage** | Azure Data Lake / Azure Blob Storage | Google Cloud Storage (GCS) | Both platforms provide scalable, cloud-based storage solutions for large datasets. |
| **Data Processing Framework** | PySpark (on Databricks) | Hadoop with MapReduce / Spark (on Dataproc) | Both environments support distributed data processing and offer Spark for faster processing. |
| **Cluster Management** | Azure Synapse Analytics / Databricks | Google Cloud Dataproc | Both offer managed services to handle distributed clusters for processing big data. |
| **Job Scheduling** | Databricks Jobs / Azure Data Factory | Cloud Composer (Airflow) / Cloud Scheduler | Both provide job orchestration and scheduling services to automate ETL workflows. |
| **Query Engine** | Synapse SQL Pool, Spark SQL | BigQuery (for querying processed data) | Both systems allow SQL-like queries over large datasets post-processing. |
| **Data Visualization** | Power BI, Azure Synapse Analytics Studio | Google Data Studio / Looker | Both have integrated visualization tools for creating dashboards and reports. |
| **Scalability** | Auto-scaling via Databricks Clusters | Auto-scaling via Dataproc | Both systems offer dynamic scaling of resources based on processing needs. |
| **Data Processing Language** | PySpark, SQL | Hadoop (MapReduce), PySpark | PySpark is available in both platforms, allowing for consistent data processing syntax. |
| **Automation** | Databricks Notebooks, Azure Logic Apps | Cloud Functions, Cloud Composer (Airflow) | Both allow automating workflows and triggering tasks based on events. |
| **Managed Services** | Databricks (fully managed) | Google Cloud Dataproc (fully managed) | Both are fully managed services for distributed data processing, abstracting infrastructure management. |