Prajyot Dhotre
SR Data Engineer
Contact: 6025706254
Email: dhotreprajyot2@gmail.com
LinkedIn:https://www.linkedin.com/in/prajyot-dhotre

Professional Summary

- Over 8 years of experience in data engineering, consistently applying deep industry knowledge and technical expertise to design, implement, and optimize complex data solutions across AWS, GCP, Azure, and various data processing technologies.
- Hands-on experience on Google Cloud Platform (GCP) in all the big data products BigQuery, Cloud Data Proc, Google Cloud Storage, and Composer (Air Flow as a service).
- Highly knowledgeable in developing data marts in the big data world in BigQuery or on-premises Hadoop clusters.
- Experience in designing and implementing scalable and effective data processing pipelines with Hadoop ecosystem components like HDFS, YARN, MapReduce, Hive, Pig, Sqoop, Spark, and Kafka.
- Managed and orchestrated containers using GKE, ensuring high availability and reliability of critical healthcare applications.
- Integrated GKE with other GCP services, such as Cloud Pub/Sub and BigQuery, to build robust and scalable data processing solutions.
- Experience with building data pipelines in Python/Pyspark/Hive SQL/Presto/BigQuery and building Python DAG in Apache Airflow.
- Proven track record in migrating on-premises Hadoop clusters to Google Cloud Platform (GCP), leveraging cloud services like BigQuery, Data Proc, and Cloud Storage for enhanced scalability, flexibility, and cost-effectiveness.
- Developed and deployed interactive dashboards and visualizations using GCP Looker, enabling stakeholders to gain real-time insights into key performance indicators and healthcare metrics.
- Proficiency in refactoring code for a seamless transition from Hadoop to GCP, ensuring compatibility and optimal performance in the cloud environment.
- Strong knowledge in data preparation, modeling, and visualization using Power BI and experience in developing various analysis services using DAX queries.
- Extensive experience in migrating and integrating data from on-premises databases and cloud sources into Snowflake, leveraging its built-in connectors and APIs.
- Hands-on experience with different programming languages such as Python, R, and SAS.
- Experience in using different Hadoop ecosystem components such as HDFS, YARN, MapReduce, Spark, Pig, Sqoop, Hive, Impala, HBase, Kafka, and Crontab tools.
- Experience in developing ETL applications on large volumes of data using different tools: MapReduce, Spark-Scala, Spark, Spark-SQL, and Pig.
- Experience in using SQOOP for importing and exporting data from RDBMS to HDFS and Hive.
- Experience with Unix/Linux systems with bash scripting experience and building data pipelines.
- Strong SQL development skills including writing Stored Procedures, Triggers, Views, and User-defined functions.
- Expertise in designing and deployment of Hadoop clusters and different Big Data analytic tools including Pig, Hive, Sqoop, and Apache Spark with Cloudera Distribution.
- Experience in building and architecting multiple Data pipelines, end-to-end ETL, and ELT processes for Data ingestion and transformation in GCP and coordinate tasks among the team.
- Hands-on experience in using the Google Cloud Platform for BigQuery, cloud data proc, and Apache airflow services. Hands-on experience on No SQL databases like HBase, Cassandra, and MongoDB.
- Experience in using stack driver service/data proc clusters in GCP for accessing logs for debugging.
- Expert knowledge and experience in dimensional modeling (Star schema, Snowflake schema), transactional modeling, and SCD (Slowly changing dimension)

Technical Skills

Big Data Technologies	HDFS, YARN, MapReduce, Hive, Pig, Impala, Sqoop, Storm, Flume, Spark, Apache Kafka, Zookeeper, Ambari, Oozie, MongoDB, Cassandra, Mahout, Puppet, Avro, Parquet, Snappy, Falcon.
NO SQL Databases	Postgres, HBase, Cassandra, MongoDB, Amazon DynamoDB, Redis
Hadoop Distributions	Cloudera (CDH3, CDH4, and CDH5), Hortonworks, MapR, and Apache.
Languages	Scala, Python, R, XML, XHTML, HTML, AJAX, CSS, SQL, PL/SQL, HiveQL, Unix, Shell Scripting
Source Code Control	GitHub, CVS, SVN, ClearCase
Cloud Services	GCP (GCP Cloud Storage, Big Query, Composer, Cloud Dataproc, Cloud SQL, Cloud Functions, Cloud Pub/Sub), Amazon AWS (S3, EMR, EC2, Lambda, VPC, Route 53, Cloud Watch, CloudFront), Microsoft Azure (Azure Storage, Azure Database, Azure Data Factory, Azure Analysis Services)
Databases	Teradata Snowflake, Microsoft SQL Server, MySQL, DB2, Oracle, PostgreSQL, and Netezza
DB languages	MySQL, PL/SQL, PostgreSQL & Oracle
Build Tools	Jenkins, Maven, Ant, Log4j
BI and Data Visualization	ETL -Informatica, SSIS, Talend, Tableau, and Power BI
Development Tools	Eclipse, IntelliJ, Microsoft SQL Studio, Toad, NetBeans
ETL Tools	Talend, Pentaho, Informatica, Ab Initio, SSIS
Development Methodologies	Agile, Scrum, Waterfall, V model, Spiral, UML
Tools	PUTTY, Putty-Gen, Eclipse, IntelliJ, and Toad
Build Tools	Apache Maven and SBT, Jenkins, Bitbucket
Operating Systems	Unix, Linux, Mac OS, CentOS, Ubuntu, and Windows

Professional Experience

Synmetta, Dallas, TX **April 2022 to Present**
Sr Data Engineer
Responsibilities:

- Played a key role in designing end-to-end architecture and migrating on-premises Oracle-based applications to Google Cloud Platform (GCP).
- Executed data ingestion, transformation, and loading into BigQuery.
- Designed and deployed serverless applications using Google Cloud Functions for event-driven data processing.
- Developed automated data validation scripts for data preservation in GCS buckets.
- Orchestrated real-time data workflows by integrating Cloud Functions with Cloud Pub/Sub.
- Built PySpark scripts for transferring data from GCP to third-party vendors through APIs.
- Developed ETL pipelines with SSIS for migrating large datasets from on-premise SQL Server to cloud storage.
- Designed scalable microservices architectures using Google Kubernetes Engine (GKE).
- Implemented CI/CD pipelines for GKE to improve deployment processes.
- Monitored and optimized GKE cluster performance and resource utilization.
- Automated backend processes like image and video processing using Cloud Functions.
- Integrated Cloud Functions with Cloud Scheduler for automating scheduled data tasks.
- Utilized Google Cloud Healthcare Data Engine for secure healthcare data ingestion and analysis.
- Designed disaster recovery and business continuity plans for GKE workloads.
- Managed and optimized GCS buckets to securely store and manage large volumes of data.

- Implemented data backup and disaster recovery strategies using GCS.
- Gathered requirements, defined data models, and implemented data integration solutions on GCP.
- Automated ETL tasks using SSIS to integrate multiple data sources into SQL Server and BigQuery.
- Designed ETL workflows using Pig, Hive, Sqoop, and Spark within the Hadoop ecosystem (Hortonworks distribution).
- Implemented data governance policies ensuring data quality, security, and compliance.
- Automated data preparation and ingestion using Qubole.
- Built interactive dashboards using Power BI for granular, actionable insights.
- Integrated Looker with BigQuery for seamless data exploration and analysis.
- Loaded streaming IoT data into BigQuery via Cloud Pub/Sub topics and Dataflow jobs.
- Developed and implemented data backup and disaster recovery strategies for critical GCP data assets.
- Configured and deployed services using GCP Cloud Shell SDK, including Cloud Dataproc, GCS, and BigQuery.
- Built analytics using SAS Visual Analytics with Oracle and Hive as data sources.
- Conducted in-depth analysis of system failures and provided corrective action recommendations.
- Integrated Looker insights into applications and workflows via Looker's API.
- Developed data pipelines in Apache Airflow (GCP Composer) using bash, Hadoop, and Python operators.
- Built automated testing frameworks for data pipelines ensuring data consistency.
- Authored Hive SQL scripts optimizing performance with partitioning, clustering, and skewing.

- Environment: GCP Console, Cloud Storage, Big Query, Data Proc, Spark, Hadoop, Hive, Scala, Cloud SQL, Shell Scripting, SQL Server 2016/2012, T-SQL, SSIS, Visual Studio, Power BI, PowerShell, Oracle, Teradata, Airflow, GIT, Docker, GCP looker

---

Allstate Corporation, Northbrook, Illinois **October 2019 to March 2022**
Data Engineer
Responsibilities:

- Designed and implemented scalable data pipelines using Google Dataflow and Apache Beam, enhancing engine performance data ingestion and analysis.
- Managed and optimized Google Cloud Storage (GCS) buckets to securely store and maintain petabytes of engine sensor data with high availability and disaster recovery.
- Automated hybrid cloud data transfers between on-premises systems and GCS using gsutil and Python scripting.
- Developed advanced SQL queries in BigQuery to support real-time analytics for engine optimization and predictive maintenance.
- Integrated Looker with Cloud Functions and Cloud Storage to automate data workflows and enhance analytical capabilities.
- Utilized Google Dataprep for cleaning, transforming, and enriching data from sources like Salesforce and Cloud Storage.
- Configured and administered GCP DataProc clusters to run large-scale Spark and Hive jobs, significantly reducing data processing time.
- Built and optimized SSIS packages for complex ETL operations across Oracle, SQL Server, and cloud databases.
- Improved BigQuery performance through the use of partitioned tables, materialized views, and query caching.
- Validated and tested Looker dashboards and data models to ensure accuracy and reliability of insights.
- Enabled object versioning in GCS for maintaining historical versions of critical data assets.

- Enhanced SSIS workflows with parallel processing and asynchronous tasks to accelerate data pipeline execution.
- Led large-scale data migration from on-premises warehouses to BigQuery using Dataflow and Transfer Service with minimal downtime.
- Implemented secure and real-time access to on-premises data sources via Power BI gateway.
- Enabled secure data sharing and collaboration across departments using Snowflake's data sharing features.
- Automated multi-service data pipeline orchestration on GCP using Cloud Composer (Apache Airflow).
- Built Snowflake ETL pipelines integrating tools like Informatica, Talend, and Apache Airflow.
- Designed and managed incremental data loading pipelines for BigQuery, Hive, Spark, and GCS using Python, Scala, Shell scripting, and gsutil.
- Deployed Cloud Pub/Sub for real-time messaging and event-driven engine performance monitoring.
- Engineered resilient, scalable storage solutions using Cloud Spanner and Cloud SQL.
- Administered and secured MySQL, PostgreSQL, and SQL Server databases on Cloud SQL, ensuring compliance and data integrity.
- Developed GCP Cloud Functions for serverless backend processing, optimizing costs and scalability.
- Leveraged GCP Data Catalog to improve metadata management, data discoverability, and governance.
- Integrated GCP Databricks for advanced analytics and machine learning, accelerating innovation in engine design.
- Managed custom data processing applications deployed on GCP VM instances for optimized resource performance.
- Developed Python and Scala applications for large-scale data processing using Spark, Hive, and Spark SQL.
- Utilized Sqoop for efficient data transfer between Hadoop and relational databases, improving ETL workflows.
- Implemented industry-standard data security, compliance frameworks, and best practices across cloud services.
- Reduced cloud infrastructure costs by carefully optimizing GCP resource provisioning and usage.
- Partnered with cross-functional teams to gather requirements and deliver comprehensive, scalable data engineering solutions.
- Provided technical leadership in migrating critical legacy systems to GCP with seamless transition and minimal disruptions.
- Built custom ETL solutions leveraging Cloud Dataflow and Apache Beam for high-volume data transformations.
- Deployed and monitored GCP resources effectively using Cloud Shell SDK and gsutil automation scripts.
- Managed BigQuery datasets efficiently using Bq command-line utilities, improving workflow automation.
- Integrated external platforms like Salesforce into GCP ecosystems to enrich analytics and business reporting.

Environment: GCP, Google Dataflow, GCP, GCS, BigQuery, GCP Dataprep, GCP Dataflow, GCP Dataproc, Cloud Composer, Cloud Pub/Sub, Cloud Storage Transfer Service, Cloud Spanner, Cloud SQL, Bucket, G-Cloud Function, Apache Beam, Cloud Dataflow, Cloud Shell, Gsutil, Bq Command Line Utilities, Dataproc, Vm Instances, Cloud Sql, Mysql, Posgres, Sql Server, Salesforce Soql, Python, Scala, Spark, Hive, Sqoop, Spark-Sql, GCP Looker

---

Docintosh Technology, Mumbai, India **February 2017 September 2019**
AWS Data Engineer
Responsibilities:

- Developed and implemented scalable and efficient data pipelines using AWS services such as S3, Glue, Kinesis, and Lambda.
- Worked with data scientists and business stakeholders to understand their requirements and design data solutions that meet their needs.
- Designed and implemented data models and data warehousing solutions using AWS services such as Redshift and Athena.
- Enhanced Power BI reports with custom visualizations and embedded analytics, allowing end-users to explore data interactively without needing technical knowledge.
- Developed and maintained ETL workflows using AWS Glue and Apache Spark.
- Built and managed streaming data pipelines using AWS Kinesis and Apache Kafka.
- Developed and implemented data processing solutions using AWS Lambda and Apache NiFi.
- Developed and implemented data security solutions using AWS services such as IAM, KMS, and S3 bucket policies.
- Worked with AWS databases such as RDS, DynamoDB, and Aurora, and implemented solutions for data replication and synchronization.
- Designed and implemented data archiving and backup solutions using AWS services such as S3 and Glacier.
- Developed and implemented data visualization solutions using AWS Quick Sight or third-party tools such as Tableau and Power BI.
- Implemented real-time data processing solutions using AWS Kinesis and AWS Lambda.
- Developed and maintained data processing workflows using Apache Airflow and AWS Glue.
- Worked with AWS machine learning services such as Sage Maker and Comprehend.
- Optimized database performance and managed ETL processes.
- Managed AWS infrastructure and resources using AWS CloudFormation or Terraform.
- Experience working with AWS VPCs, subnets, security groups, and load balancers.
- Knowledge of AWS networking concepts and experience implementing and managing AWS Direct Connect, VPN, and Route53.
- Performed Hive test queries on local sample files and HDFS files.
- Used Spark Streaming to divide streaming data into batches as an input to the spark engine for batch processing.
- Worked on analyzing Hadoop cluster and different Big Data analytic tools including Pig, hive, HBase, Spark, and Sqoop.
- Generating various capacity planning reports (graphical) using Python packages like NumPy, and matplotlib.
- Analyzing various logs that are been generating and predicting/forecasting the next occurrence of an event with various Python libraries.
- ETL pipelines in and out of the data warehouse using a combination of Python and Snowflakes Snow SQL Writing SQL queries against Snowflake.

Environment: AWS (EC2, S3, EBS, ELB, RDS, SNS, SQS, VPC, LAM Cloud formation, CloudWatch, ELK Stack), Bitbucket, Ansible, Python, Shell Scripting, PowerShell, GIT, Jira, JBOSS, Bamboo, Docker, Web Logic, Maven, Web sphere, Unix/Linux, AWS X-ray, Dynamodb, Kinesis, CodeDeploy, CodePieline, CodeBuild, CodeCommit, Splunk, SonarQube

---

Education
Bachelor of Science in Computer Science (2017)
Mumbai University, Mumbai, India