DATA ENGINEERING TOOLS – SEGREGATED

Segment	General	AWS
Data Ingestion	**Ingestion Tools:** Apache Kafka Apache NiFi AWS Kinesis Logstash	**AWS Glue:** Use Glue Crawlers to discover and catalog metadata from various data sources. Glue ETL jobs for transforming and loading data. **Amazon Kinesis:** Kinesis Data Streams for real-time data streaming. Kinesis Data Firehose for loading streaming data into data stores. **AWS DataSync:** Transfer data from on-premises to AWS.
Data Storage	**Data Warehouses:** Amazon Redshift Google BigQuery Snowflake **Data Lakes:** Amazon S3 Azure Data Lake Storage Google Cloud Storage **Databases:** PostgreSQL MySQL MongoDB Cassandra	**Amazon S3:** As a data lake for storing raw and processed data. Versioning and lifecycle policies for managing data. **Amazon Redshift:** For data warehousing and complex queries. **Amazon DynamoDB:** For NoSQL database requirements.
Data Processing	**Batch Processing:** Apache Spark Apache Flink Hadoop MapReduce **Stream Processing:** Apache Kafka Streams Apache Storm Apache Flink **ETL (Extract,Transform, Load):** Apache Beam Apache Airflow Talend	**Amazon EMR (Elastic MapReduce):**For big data processing using frameworks like Apache Spark and Hadoop. **AWS Glue:** Serverless ETL service for data transformation and preparation. **AWS Lambda:** For serverless event-driven processing.
Data Transformation	**Data Preparation:** Pandas (Python library) Apache Beam **Data Cleansing:** Trifacta OpenRefine **Data Masking/Anonymization:** Google DLP Apache Nifi	**AWS Glue:** Use Glue jobs for ETL transformations. **AWS Step Functions:** Orchestrate and coordinate multiple AWS services in a serverless workflow.

Analytics and Reporting	**Business Intelligence Tools:** Tableau Power BI Looker **Analytics Platforms:** Databricks Google Analytics Mixpanel	**Amazon QuickSight:** Business intelligence service for visualizing and analyzing data. **Amazon Athena:** Serverless query service for analyzing data in Amazon S3.
Data Orchestration	**Workflow Management:** Apache Airflow Luigi Prefect **Job Scheduling:** Cron Apache Oozie	**Apache Airflow on Amazon MWAA (Managed Workflows for Apache Airflow):** Orchestrate and schedule complex data workflows. **AWS Step Functions:** For serverless workflow orchestration.
Monitoring and Logging	**Logging:** ELK Stack (Elasticsearch, Logstash, Kibana) Splunk **Monitoring:** Prometheus Grafana	**Amazon CloudWatch:** For monitoring AWS resources and applications. **AWS CloudTrail:** For logging AWS API calls.
Data Data Quality and Governance	**Data Quality Tools:** Informatica Talend Apache Griffin **Metadata Management:** Collibra Apache Atlas	**AWS Glue DataBrew:** For data profiling, cleaning, and exploration. **AWS Lake Formation:** Set up and enforce security, governance, and auditing policies.
Security and Access Control	**Encryption:** TLS/SSL HDFS Encryption **Access Control:** Apache Ranger AWS IAM Google Cloud Identity and Access Management (IAM)	**AWS IAM (Identity and Access Management):** Manage access to AWS resources. **AWS Key Management Service (KMS):** Encrypt data at rest and in transit.
Data Science Integration	**Model Deployment:** TensorFlow Serving MLflow PMML (Predictive Model Markup Language) **Notebook Environments:** Jupyter Notebooks Google Colab Databricks Notebooks	**Amazon SageMaker:** For building, training, and deploying machine learning models.
Architectural Patterns	**Lambda Architecture:** Combines batch and stream processing for real-time and batch processing.	**Serverless Architecture:** Leverage services like Lambda, Glue, and Step Functions for serverless processing.

	Kappa Architecture: Simplifies the Lambda Architecture using only stream processing.	**Data Lake Architecture:** Utilize S3 as a central data lake to store structured and unstructured data.
Data Versioning and Lineage	**Version Control:** Git DVC (Data Version Control) **Lineage Tracking:** Apache Atlas DataHub	
Cloud Integration	**Cloud Platforms:** AWS, Azure, Google Cloud Platform (GCP) **Serverless Computing:** AWS Lambda Azure Functions Google Cloud Functions	**AWS Direct Connect or VPN:** Connect on-premises data centers to AWS. **AWS SDKs and CLI:** Integrate and automate AWS services using SDKs and the Command Line Interface.

Segment	Microsoft Azure	Google Cloud Platform
Data Ingestion	**Azure Data Factory:** Orchestrate and automate data workflows. Support for data movement from various sources to data lakes or warehouses. **Azure Event Hubs:** Ingest and process massive amounts of streaming data.	**Cloud Pub/Sub:** Real-time messaging service for event-driven architectures. **Cloud Storage:** Object storage for batch uploads.
Data Storage	**Azure Data Lake Storage:** Scalable and secure data lake storage. **Azure SQL Data Warehouse (now part of Azure Synapse Analytics):** Enterprise-grade analytics service. **Azure Cosmos DB:** Globally distributed, multi-model database for operational and analytical workloads.	**BigQuery:** Fully-managed, serverless data warehouse for analytics. **Cloud Storage:** Object storage for raw data and backups. **CloudSQL:** Managed relational databases.
Data Processing	**Azure Databricks:** Apache Spark-based analytics platform for big data and machine learning. **HDInsight:** Fully managed cloud service for big data analytics using Hadoop, Spark, HBase, and more. **Azure Stream Analytics:** Real-time analytics on streaming data.	**Dataflow:** Fully managed stream and batch processing using Apache Beam. **Dataprep by Trifacta:** Cloud-native data preparation service. **Dataproc:** Managed Apache Spark and Hadoop service.
Data Transformation	**Azure Data Factory:** Transform and clean data using data flows and transformations. **Azure HDInsight:** Leverage Apache Spark or Hive for data transformation.	**Dataflow:** Apache Beam for ETL pipelines. **Cloud Dataprep:** Visual data preparation tool.

Analytics and Reporting	**Power BI:** Business Intelligence and visualization. **Azure Synapse Studio:** Integrated analytics and data exploration.	**BigQuery:** For ad-hoc queries and analytics. **Looker, Tableau, or Data Studio:** Business intelligence and visualization tools.
Data Orchestration	**Azure Data Factory:** Schedule and orchestrate data workflows. **Azure Logic Apps:** Automate workflows and integrate services, including data services.	**Cloud Composer:** Managed Apache Airflow for workflow orchestration. **Cloud Scheduler:** Fully managed cron job scheduler.
Monitoring and Logging	**Azure Monitor:** Monitor the performance and health of resources. **Azure Log Analytics:** Collect and analyze log data.	**Cloud Monitoring:** Infrastructure and application monitoring. **Cloud Logging:** Centralized log management.
Data Data Quality and Governance	**Azure Purview:** Unified data governance service for discovering, understanding, and managing data. **Azure Data Catalog:** Discover, register, and manage data asset.	**Cloud Data Catalog:** Fully managed and scalable metadata management service. **Cloud Data Loss Prevention (DLP):** Sensitive data discovery and redaction.
Security and Access Control	**Azure Active Directory (AAD):** Identity and access management. **Azure Key Vault:** Securely store and manage sensitive information like keys and secrets.	**Cloud Identity and Access Management (IAM):** Access control for GCP resources. **Cloud Key Management Service (KMS):** Manage cryptographic keys.
Data Science Integration	**Azure Machine Learning:** End-to-end platform for building, training, and deploying machine learning models.	**AI Platform:** Managed services for building, training, and deploying machine learning models. **Notebooks:** AI Platform Notebooks or Jupyter Notebooks on AI Platform.
Architectural Patterns	**Modern Data Warehouse (Azure Synapse Analytics):** Combines big data and data warehousing for analytics. **Event-Driven Architectures:** Use Azure Event Hubs and Azure Functions for event-driven processing.	**Serverless Architecture:** Utilize serverless services like Cloud Functions. **Data Lake and Data Warehouse:** Combine Cloud Storage and BigQuery for cost-effective storage and analytics.
Data Versioning and Lineage		**Cloud Data Catalog:** Track and manage data lineage. **BigQuery:** Keep track of changes with versioned tables.
Cloud Integration	**Azure Functions:** Serverless computing for event-driven solutions. **Azure Logic Apps:** Connect and automate workflows across cloud and on-premises services.	**Cloud Functions:** Serverless computing for event-driven functions. **Cloud Run:** Fully managed compute platform for containerized applications.