Data lake

The main components of Data Late are:

— Data Ingestion

Data Ingestion is the process of collecting data from a variety of data sources and load it into the lake. A well-architected ingestion layer should support structured, semi-structured, unstructured data from multiple data sources like Iot, Emails, Social Media, Databases, FTP etc. Apart from these, it should be flexible enough to support new data sources. It should also have facilities to curate the data.

— Data Storage

Any Data Lake architecture should have data storage capability. A well-architected storage layer should be able to store curated, raw, in-process data of any format with compression and encryption techniques. More than that, it should be low-cost, highly available and scalable, and it should provide fast access for data manipulation.

— Data Governance

Data Governance involve Scheduling, Monitoring, Provisioning, Managing Hadoop clusters.

- Security: It involves authenticating, authorizing, accounting and protecting data by restricting the access to trusted users and serves only.

- Quality: Data quality should be maintained to bring business value out of the lake. It involves Discovering data, monitoring Data Quality, profiling data, and defining data Quality Rules.

- Data Lineage: It keep an eye on the data and traces what happens to the data and where the data moves over the time.

- Data Auditing: It involves tracking changes to key dataset elements and capture "who / when / how" information about changes to these elements

— Data Discovery

It is the process of understanding data and tagging through identifying, organizing, and interpreting the raw data ingested in the lake.

— Data Exploration and Visualization

To explore and visualize, the real consumers should have the ability to flexibly access the data through friendly Dashboards and explore it independent of IT.

Comparison

	Azure Stream Analytics	*AWS Kinesis Data Analytics*	*Google Cloud Dataflow*	*IBM Streaming Analytics*	*Alibaba Data Lake Analytics*
Programmability	Stream analytics query language, JavaScript	Data analytics query language, standard SQL	Java, Python and a distributed compute platform	Java, Scala and Python	Standard SQL
Programming model	Declarative	Flink programming model, Declarative	Apache Beam	Streams Processing Language (SPL), Declarative	Declarative
Pricing model	Streaming units	Hourly rate based on the average streaming units	Based on Google Compute Engine (GCE) costs plus an additional charge per vCPU per minute	Subscription based	Based on the number of bytes scanned
Inputs	Azure Event Hubs, Azure IoT Hub, Azure Blob storage	Data sources through SQL JOINS: Streaming data sources like Kinesis Data Streams and reference data sources like Amazon S3	Cloud Storage and PubSub	File, Transmission Control Protocol (TCP) and User Datagram Protocol (UDP)	Alibaba Cloud Object Storage Service (OSS), PostgreSQL, MySQL, NoSQL (Table Store) and ApsaraDB, using DLA and Quick BI
Sinks	Azure Data Lake Store, Azure SQL Database, Storage Blobs, Event Hubs, Power BI, Table Storage, Service Bus Queues, Service Bus Topics, Cosmos DB, Azure Functions	Amazon Kinesis Data Streams, Amazon Kinesis Data Firehose, Amazon DynamoDB, and Amazon S3 (through file sink integrations)	Cloud Storage, BigQuery, BigTable, PubSub, Datastore, etc.	TCP network connection, UDP network connection and User-defined Sink Operator	NA
Built-in temporal/windowing support	Yes	Yes	NA	Yes	NA
Input data formats	Avro, JSON or CSV, UTF-8 encoded	JSON, CSV, and TSV	AVRO, CSV, JSON	JSON	JSON, Vector and other multi-media resources
Scalability	Query partitions	Shards	Shards	Horizontal partitions	Horizontal partitions
Late arrival and out of order event handling support	Yes	Yes	Yes	NA	Yes

Operating systems supported:

Area	**AWS**	**GCP**	**IBM Cloud**	**Azure**	**Alibaba Cloud**
Windows	Yes	Yes	Yes	Yes	Yes
SLES	Yes	Yes	No	Yes	Yes
CentOS	Yes	Yes	Yes	Yes	Yes
CoreOS	Yes	Yes	Yes	Yes	Yes
RHEL	Yes	Yes	Yes	Yes	Yes
CloudLinux	Yes	No	Yes	No	No
OpenSUSE	Yes	Yes	NO	Yes	Yes
FreeBSD	Yes	Yes	Yes	Yes	Yes
Ubuntu	Yes	Yes	Yes	Yes	Yes
Debian	Yes	Yes	Yes	Yes	Yes
Vyatta	Yes	No	Yes	No	No
Oracle Linux	Yes	No	No	Yes	No

Compute:

Area	**AWS**	**GCP**	**IBM Cloud**	**Azure**	**Alibaba Cloud**
Scalability	AWS Auto Scaling	Autoscaling	Auto Scaling	Azure Autoscaling Virtual Machine Scale Sets App Service Scale Capability	Autoscaling
Virtual Servers	Elastic Compute Cloud (EC2) Instances	Custom Machine Types	IBM Virtual Servers	Azure Virtual Machines	Simple Application Server

	Amazon LightSail	Compute Engine		Virtual Machine Images	Elastic Compute Service
Container Instances	EC2 Container Service (ECS)	Kubernetes Engine	IBM Cloud Kubernetes Service	Azure Kubernetes Service (AKS)	Container Registry
	Elastic Container Registry (ECR)			Azure Container Registry	
Container Orchestrators/ Microservices	Elastic Container Service for Kubernetes (EKS)	Kubernetes Engine	IBM Cloud Kubernetes Service	Azure Kubernetes Service (AKS)	Container Service
				Service Fabric	Container Service for Kubernetes
Job Orchestration	AWS Batch	Preemptible VMs	No	Azure Batch	Batch Compute
Serverless Computing	Lambda	Google Cloud Functions	IBM Cloud Functions	Azure Functions	Function Computes
	Lambda @ Edge			Azure Event Grid	
Time sync	Time Sync Service	TrueTime Service	IBM App Connect	Time sync	Time Setting

Storage:

Area	AWS	GCP	IBM Cloud	Azure	Alibaba Cloud
Object storage	Simple Storage Services (S3)	Google Cloud Storage	IBM Cloud Object Storage	Azure Blob Storage	Object Storage Service
Shared file storage	Elastic File System	Google Cloud Storage FUSE	File Storage	Azure Files	Network Attached Storage
Virtual Server disk infrastructure	Elastic Block Store (EBS)	Google Persistent Disk	Block Storage	Disk Storage – Premium Storage	Block Storage

				Disk Storage – Page Blobs (for VHDs or other random-write type data)	
Archiving – cool storage	S3 Standard Infrequent Access (IA)	Google Nearline Storage	Object Storage	Storage –Hot, Cool & Archive Tier	Object Storage Service
Archiving – cold storage	S3 Glacier	Google Cloud Storage (GCS) Coldline	Backup Storage	No	Object Storage Service
Hybrid Storage	Storage Gateway	No	No	StorSimple	Hybrid Cloud Storage Array
Backup	AWS Cloud Backup	No	IBM Cloud Backup	Azure Backup	Cloud Backup and Recovery Database Backup
Data transfer	AWS Import/Export Disk	Cloud Data Transfer Appliance	Data Transfer Service	Azure Import/ Export	Data Transport
	AWS Import/Export Snowball		Mass Data Migration		
	AWS Snowmobile			Azure Data Box	Cloud Migration Tool
	AWS Snowball Edge				
Disaster Recovery	CloudEndure Disaster Recovery	No	No	Site Recovery	Alibaba Disaster Recovery Hybrid Backup Recovery

Database Supported:

Area	AWS	GCP	IBM Cloud	Azure	Alibaba Cloud

Relational Database	RDS for MariaDB	SQL Server	Compose for MySQL	SQL Database	ApsaraDB for MySQL
	RDS for SQL Server	Google Cloud SQL	Compose for Postgre SQL	Azure Database for MySQL	ApsaraDB for MariaDB TX
	RDS for MySQL	Cloud SQL support for Postgre SQL (Beta)	IBM Cloud Databases for EnterpriseDB	Azure Database for PostgreSQL (Preview)	ApsaraDB RDS for SQL server
	RDS for Oracle DB			Azure Database for MariaDB	ApsaraDB for PPAS
	RDS for Postgre SQL				
		Cloud Spanner	Db2 on Cloud		ApsaraDB for PostgreSQL
					Distributed Relational Database Service (DRDS)
					ApsaraDB for POLARDB
NoSQL – key/value storage, document storage	Dynamo DB		Database for Redis	Table Storage	Time Series Database
			Databases for MongoDB Databases for Elasticsearch	Azure Cosmos DB	
	SimpleDB				
Non-relational Databases	Amazon Neptune	Cloud BigTable	Compose for JunusGraph	Azure HDInsight	ApsaraDB for Redis
	Amazon EMR	Cloud Firestore		Azure Batch	ApsaraDB for MongoDB
	Amazon Dynamo DB	Firebase Realtime database			
	Amazon SimpleDB	Cloud Memorystore	Cloudant	Cosmos DB	ApsaraDB for Memcache
Database Migration	Database Migration Service	Cloud Data Transfer	Lift	Azure Database Migration Service	Data Transmission Services

				Data Migration Assistant	
Caching	ElastiCache	Memorystore	Compose for Redis	Azure Redis Cache	ApsaraDB for Redis

Analytics and Big data:

Area	AWS	GCP	IBM Cloud	Azure	Alibaba Cloud
Elastic data warehouse	Amazon Redshift	Google Cloud BigQuery	Db2 Warehouse on Cloud	SQL Data Warehouse	Max Compute
Data orchestration	Data Pipeline AWS Glue	Google Cloud Dataflow	No	Data Factory Data Catalog	DataWorks Data Integration
Big data processing	Elastic MapReduce (EMR)	Google Cloud Dataproc Dataflow	IBM Analytics Engine	HDInsight	E-MapReduce Realtime compute
Data discovery	Amazon Athena	Google BigQuery	SQL Query	Data Catalog Azure Data Lake Analytics	DataWorks
Search	Amazon Elasticsearch	No	Compose for Elasticsearch	Azure Search	Alibaba Cloud Elasticsearch

	Amazon CloudSearch			Marketplace – Elasticsearch	
Analytics	Kinesis Data Analytics	Google Cloud Dataflow	Streaming analytics	Stream Analytics	AnalyticDB
				Data Lake Analytics	Alibaba Cloud Elasticsearch
				Data Lake Store	
Visualization	Amazon QuickSight	Google Data Studio	No	PowerBI	DataWorks
				PowerBI Embedded	DataV
Machine Learning	Machine Learning	Google Cloud AI	Watson Machine Learning	Azure Machine Learning Studio	Machine Learning Platform for AI
	SageMaker	Google Cloud Datalab		Azure Machine Learning Workbench	
		Google Cloud Machine Learning Engine			

Identity, Access and security:

Area	AWS	GCP	IBM Cloud	Azure	Alibaba Cloud
Firewall	Web Application Firewall	Google Cloud Firewall	Firewalls	Application Gateway	Application Gateway

	AWS Firewall Manager			Web Application Firewall (in preview)	
Authorization & Authentication	Identity and Access Management (IAM) AWS Single Sign-On	Google Cloud Identity and Access Management	IBM Cloud App ID	Azure Active Directory	Resource Access Management
	Multi-Factor Authentication			Azure Subscription and Service Management + RBAC	
	AWS Organizations			Multi-Factor Authentication	
Encryption	Amazon S3 Key Management Service for server-side encryption	Google Cloud Key Management Service	Hardware Security Module IBM Hyper Protect Crypto Services IBM Cloud Data Shield	Azure Key Vault	Key Management Service
	Key Management Service		Key Protect	Storage Service Encryption	
	CloudHSM				
Security assessment and threat detection	Amazon Inspector	Security Command Center Security Health Analytics	SSL Certificates	Security Center	Alibaba Cloud SSL Certificates Service
	Guard Duty		No		
	Certificate Manager		Nessus Security Scanner	App Service Certificates on the portal	
Compliance	AWS Artifact	Google Cloud Compliance	IBM Cloud Compliance	Microsoft Service Trust Portal	Security & Compliance Center

Directory Services	AWS Directory Service + Windows Server Active Directory on AWS	Google Cloud Directory Sync	No	Azure Directory Domain Services + Windows Server Active Directory on Azure IaaS	Resource Access Management
	Cognito			Azure Active Directory B2C	
	AWS Directory Service				
Information protection	AWS Cloud Security	No	No	Azure Information Protection	Managed Security Service

Monitoring and Management:

Area	AWS	GCP	IBM Cloud	Azure	Alibaba Cloud
DevOps (Deployment Orchestration)	OpsWorks	Cloud Composer	IBM Cloud Deployment Services	Azure Automation	DevOps Solution
	CloudFormation			Azure Resource Manager	Resource Orchestration Service
				VM extensions	
Monitoring & Management (DevOps)	Amazon CloudWatch	Google Operations Logging, Monitoring	IT Operations Management	Azure portal	Cloud Monitor
	AWS CloudTrail	Debugger		Azure Monitor	Log Service
	AWS X-Ray	Error Reporting	IT Operations Analytics	Azure Application Insights	ActionTrail
	AWS Cost and Usage Report	Google Trace		Azure Billing API	Application Real-time monitoring service
	AWS Management Console			Cloud Shell	
				Log Analytics	

Cloud advisor	Trusted Advisor	Consulting Services	Advisory Services	Azure Advisor	Professional Services
Administration	AWS Application Discovery Service	Cloud Console	IBM Cloud Orchestrator	Azure Log Analytics in Operations Management Suite	Application Real-time monitoring service
	AWS Systems Manager			Microsoft Operations Management Suite – Automation and Control functionalities	
	AWS Personal Health Dashboard			Azure Resource Health	
				Azure Storage Explorer	