|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Company** | **Data Lake Products & Links** | **Features** | **Company Data Lake Products Pros & Cons** | **Why to choose this Data Lake Platform over the other companies?** |
| Amazon Web Services | **Amazon S3**  **Links: https://aws.amazon.com/s3/?c=a&sec=srvm** | \* Object Storage Service  \* Append Metadata to objects  \* Secure data with encryption features  \* Blocks public access  \*S3 Storage Classes – Support different data access levels & retains objects long-term at the lowest rates  \*S3 Storage Class Analysis – Identify data that should be moved down to a lower-cost storage based on access patterns  \*S3 Object Lambda – Add your own code to process data retrieved from S3  \* Manage data access with specific permissions using shared data set  \*Customers can create DR architectures in order to easily recover from outages  \* Can upload any amount of data and access the data anywhere | **Pros:**  \*Dominance of public cloud market for the last 10 years  \* Offers huge and growing array of available services with the most comprehensive network for worldwide centers  \* Extensive Training  \* Global reach  **Cons:**  \* Highest cost when compared to Microsoft and Google  \* Cannot offer a personal relationship | AWS is the number one provider for a reason. AWS offers plenty features to accommodate all different kinds of business whether a small business to big enterprise. With AWS being serverless, the business doesn’t have to worry about infrastructure, management AWS will handle all that for the company. This allows to focus on more pressing matters. AWS offers reaches worldwide and still offers their customers good support and extensive training. Although pricing can be steep for AWS, however it’s worth the money when it comes to the products and services received. |
| **Amazon S3 Glacier**  **https://aws.amazon.com/glacier/?c=a&sec=srvm** | \*Back up and archive  \* Long-term, secure object storage classes for data archiving starting at $1 per terabyte per month  \*Expedited: Retrievals data as quick as 1-5 Minutes  \* Standard: 3-5 hours  \*Bulk: 5-12 hours  \*Data is automatically distributed across a min of three physical availability zones  \*Offers integration with AWS CloudTrail to log, monitor and retain storage API call activates for auditing.  \*Low Cost  \*Supported by partners, vendors and AWS Services. |
| **AWS Glue**  **https://aws.amazon.com/glue/?c=a&sec=srvm&whats-new-cards.sort-by=item.additionalFields.postDateTime&whats-new-cards.sort-order=desc** | \*Data catalog  \*Serverless  \* Visual and code-based interfaces to make data integration easier  \* Build event driven ETL pipelines  \* Create a unified catalog to find data across multiple data stores  \* Create, run, and monitor ETL jobs without coding  \* Explore and experiment with data directly from your data lake, data warehouses, and databases, including Amazon S3, Amazon Redshift, AWS Lake Formation, Amazon Aurora, and Amazon RDS.  \* Self-service visual data preparation  \* Build materialized views to combine and replicate data  \* Deduplicate and cleanse data with built-in machine learning  \* Pricing: Pay an hourly rate, billed by the second, for crawlers (discovering data) and ETL jobs (processing and loading data). For the AWS Glue Data Catalog, you pay a simple monthly fee for storing and accessing the metadata. The first million objects stored are free, and the first million accesses are free. |
| **AWS Lake Formation**  **https://aws.amazon.com/lake-formation/?whats-new-cards.sort-by=item.additionalFields.postDateTime&whats-new-cards.sort-order=desc** | \* Centralized, curated, and secured repository  \* Build data lakes quickly  \*Eliminates data silos  \*Helps collect and catalog data from databases  \* Automatically changes data into formats like Apache Parquet and ORC for faster analytics.  \*Built in machine learning  \*Centralized security, governance and auditing  \*Provides Self-Service access to data |
| **AWS Data Exchange**  **https://aws.amazon.com/data-exchange/?c=a&sec=srvm** | \*Third-party data  \* Quickly find diverse data in one place  \* Efficiently access data in the cloud  \*Loads directly into Amazon S3.  \* Removes the need to build and maintain infrastructure  \* Has over 1000 products that can be used across multiple industries.  \* Offers trails before users have to subscribe  \* Eliminates the need to receive physical media, manage FTP credentials or APIs from multiple providers |
| **Google Cloud** | **Big Query**  **https://cloud.google.com/bigquery** | \*Serverless  \*New customers get $300 in free credits to spend on Google Cloud during the first 90 days. All customers get 10 GB storage and up to 1 TB queries/month, completely free of charge.  \*Built-in machine learning  \*Business decisions from data across clouds  \*Analyze petabytes of data using ANSI SQL at blazing-fast speeds  \*Real-time and predictive analytics  \*Robust security, governance, and reliability controls that offer high availability and a 99.99% uptime SLA.  \* Build and operationalize ML models on planet-scale structured or semi-structured data by using simple SQL in fraction of time.  \* In-memory analysis service built into BigQuery that enables users to analyze large and complex datasets  \*Allows users to analyze billions of rows of live BigQuery data in Google Sheets without requiring SQL knowledge. | **Pros:**  \*Strong Offering in containers  \* Specializes in high compute offerings like Big Data, analytics and machine learning  \* Fast response times  \* Supports open source  **Cons:**  \* Ranked 3rd in the market when being compared to Microsoft & AWS  \* Fewer features  \*Designed for cloud native businesses  \*Lacks feature  \* Difficulty migrating and deploying Google cloud. | Google continues to improve over the years by offering more products at reasonable pricing. Google cloud customers are offered a free tier that does not expire. Additionally, new customers get 300 credits so customers can explore more options Google offer. On top of that google offers a wide variety of security products to keep your data secure. Google security meets compliance regulations of CSA STAR, PCI DSS, and ISO/IEC. |
| **Dataflow**  [**https://cloud.google.com/dataflow/**](https://cloud.google.com/dataflow/) | \*Fully managed data processing service  \*Automated provision and management of processing resources  \*Horizontal autoscaling of worker resources to maximize resource utilization  \*Reliable and consistent exactly–once processing  \*Streaming data analytics with speed  \*Serverless  \*Flexible scheduling and prices for batch processing  \*Ready to use AI  \*Dataflow shuffle  \*Data SQL |
| **Dataproc**  **ttps://cloud.google.com/dataproc** | \*Scalable service for running Apache Spark, Apache Flink, Presto, and 30+ open-source tools and frameworks.  \*Build fully managed open-source clusters on demand, integrated with [data lake metastore](https://cloud.google.com/dataproc-metastore/docs?skip_cache=true)  \*Run clusters in [high availability mode](https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/high-availability) with multiple master nodes and set jobs to restart on failure t  \* Reduce TCO of Apache Spark management by [up to 57%](https://www.esg-global.com/analyzing-the-economic-benefits-of-google-cloud-dataproc).  \*Containerize Apache Spark jobs with Kubernetes  \*Enterprise security integrated with Google Cloud  \*Built-in integration with Cloud Storage, BigQuery, Cloud Bigtable, Cloud Logging, Cloud Monitoring, and AI Hub, giving you a more complete and robust data platform.  \*Cluster scheduled deletion to avoid incurring charges for inactive cluster  \*Automatic or manual configuration while giving users manual control  \*Developer tools to help manage east to use web UI, cloud SDK, RESTful APIs, SSH access |
| **Cloud Storage**  **https://cloud.google.com/storage** | \*Cloud storage supports high-volume ingestion of new data and high-volume consumption of stored data.  \*Designed for 99.999999999% annual durability  \* Provides support for read-after-write operations, listing buckets, and objects and granting access to resources  \*Cost Effective – Provides multiple prices to suit different access patterns and availability needs.  \*Provides integration with BigQuery, Dataproc, Dataflow, Video Intelligence API and cloud vision to give customers flexibility to choose the right tool.  \*Central Repository  \*Cloud storage offers a number of mechanisms to implement so users can control access to data.  \* Unlimited storage with no min object size  \*Worldwide accessibility  \*Low latency |
| **Cloud Data Fusion**  **https://cloud.google.com/data-fusion** | \*Built using open-sourced project CDAP ensures data pipeline portability for users. \* Integrated with Google’s industry-leading big data tools which simplifies security and ensures data is available. \* Offers pre-built transformations for both batch and real time processing  \* Code free self-service  \* Users create an internal library of custom connections and transformations that can validate, share, and reuse across an organization.  \*Real-time data integration  \* Batch integration  \* Enterprise-grade security  \*Integration metadata and lineage  \*REST APIs, time-based schedules, pipeline state-based triggers, logs, metrics, and monitoring dashboards |
| **Microsoft Azure** | **Azure Data Lake Storage**  **https://azure.microsoft.com/en-us/services/storage/data-lake-storage/** | \*Optimize costs with tiered storage and policy management.  \*Protect data with security features like encryption at rest and advanced threat protection.  \*Limitless scale and 16 9s of data durability with automatic geo-replication  \*Single storage platform for ingestion, processing, and visualization that supports the most common analytics frameworks  \*Process data using Azure Databricks, Synapse Analytics, or HDInsight. And visualize the data with Microsoft Power BI for transformational insights  \*Run large-scale analytics queries at consistently high performance | **Pros:**  **\***Ranked second largest provider  **\***Repurposed Windows Service, Office, SQL Server, SharePoint, Dynamics Active Directory, .NET, and other and repurposing it for the cloud  \*Enterprise ready  \*Price – If an existing Microsoft customer you can expect significant discounts**.**  **\*** Hybrid Cloud  \*Supports Open source  **Cons:**  **\***Poor Customer support  \* Only subscription management and billing support included in Azure subscription.  **\***Technical support is only provided with plans starting at 129 to 1000.  **\***Outages with sometimes long recovery times. | Microsoft Azure is used by 95% of fortune 500 companies. Azure offers o customers over 200 cloud products with the most comprehensive set of compliance offering. Azure has a security center that combines security management and advance threat production across all hybrid clouds. Azure will automatically scale with the changing demand in order to save customers money. |
| **Azure HDInsight**  **https://azure.microsoft.com/en-us/services/hdinsight/** | \*Pay for what you need  \*Flexibility of multiple languages and tools such as visual studio, Eclipse, Python, R, JavaScript, and etc.  \*Integration with Azure data storage solutions and services including Azure Synapse Analytics, Azure Cosmos DB, Azure Data Lake Storage, Azure Blob Storage, Azure Event Hubs, and Azure Data Factory  \*Build your projects in an open-source ecosystem  \*Enterprise-grade security and industry-leading compliance with more than 30 certifications helps protect your data  \*Optimized components for open-source technologies such as Hadoop and Spark keep you up to date  \*Gives you full control over the configuration of your clusters and the software installed on them.  \* Can process historical data, or real time data |
| **Data Lake Analytics** | \*On demand analytics job service that simplifies big data  \*No infrastructure to manage, you can process data on demand, scale instantly, and only pay per job.  \*U-SQL allows you to write code once and have it automatically parallelized for the scale you need  \*Execution environment actively analyzes your programs as they run and gives you recommendations to improve performance and reduce cost.  \*Queries are automatically optimized by moving processing close to the source data without data movement, which maximizes performance and minimizes latency.  \*Guarantee a 99.9% enterprise-grade SLA and 24/7 support for your big data solution  \*Single sign-on (SSO), multi-factor authentication, and seamless management of millions of identities are built-in through Azure Active Directory.  \* Pricing – No upfront cost, no termination fees, pay only for what you use and per second billing |