Cloud based Architecture

A sample architecture design is available parallel to this file

Components/services:-

- CI- engine (data producers)
- Amazon Kinesis data firehose
- Amazon Kinesis data analytics
- AWS Lambda functions
- AWS S3 data_lake
- Amazon Redshift
- Amazon quick-sight

Design description:-

- Ci –engines (Jenkins) will be generating the logs based on daily builds and various metrics will be captured like build status, average runtime of builds and etc.
- Amazon kinesis agent will configured on the build servers where it can read the log files and send data to AWS kinesis firehose
- Kinesis agent is a standalone java application which offers an easy way of collecting data and transferring.
- It continuously monitors the set of files and send data to firehose delivery Strems.
- Agent only handles the file rotation, check pointing, and retry upon failures.
- Once the data is at firehose it will streamline data to data to Aws kinesis data analytics for transforming and analyzing the data in real time using apache Flink framework and engine for processing data streams.
- After processing the data, its transferred to a kinesis firehose for sending it to final destinations

- An AWS lambda function can be used to further transform the data for final storage in the S3 buckets
- And it can transfer to amazon redshift for data visualization.
- Amazon quick_search service will be used to query and visualize data for monitoring and analysis

Service features:-

CI-engines:-

- We can have multiple Ci-engines for data generation and which can be easily handled for kinesis firehose with less configuration changes and efforts

Amazon Kinesis data firehose:-

- It's a reliable service for pipelining data for source to destinations
- It can handle multiple sources and multiple destinations seamlessly.
- Data collections for sources can be done in various ways
- We can use kinesis agent which can be configured in the build servers for collecting data
- We can even use the firehose API for the collection of data
- Data delivery to the destinations are done using kinesis delivery streams
- We can have multiple data sources and delivery streams.
- It offers a real-time data movement.

Benefits of kinesis firehose:-

- data is processed synchronously
- 2) data records can be as high as 1000KB
- 3) it can transfer to diff destinations like s3 buckets, redshift clusters, elastic search clusters
- 4) its highly reliable and secure

- 5) data is encrypted using AWS Key Management Service (AWS KMS) key before storing the data to s3 buckets

Amazon Kinesis data Analytics:-

- It offers easiest way to transform and analyze streaming data in real time with apache Flink.
- Apache Flink is an open source framework and engine for processing data streams
- It automatically scales to make the throughput of incoming data
- It is a server less feature so no sever maintenance required
- No setup cost required and it offers pay as how much data is streamed.

Benefits of Kinesis data analytics:-

- It has built-in functions to filter and transform incoming data streams.
- As its offering real-time analysis, we can respond to events in real time.
- If we are not satisfied with inbuilt functions we can implement AWS lambda functions for the data processing.

AWS lambda functions:-

- It can be used and alternative to kinesis data analytics for implementing custom logics.
- It's a sever less service and runs in real time in response to incoming triggers.
- It's a highly reliable and cost effective.
- It's highly scalable and it can rapidly launch as many copies of the function as needed to scale to the rate of incoming events.
- We can associate lambda functions to various other services and it will be triggered based on the change events.

Benefits of Lambda functions:-

- Bring your own code, we can implement any functionality written in any languages, it supports Java, Go, PowerShell, Node.js, C#, Python, and Ruby code etc.
- It has built-in fault tolerance as its compute capacity spreads across multiple availability zones and regions.
- It offers auto scaling.
- We can co-ordinate multiple lambda functions to implement complex or long running tasks.
- Lambda allows your code to securely access other AWS services through AWS –IAM

Amazon S3 (simple storage service) data lake:-

- It's a highly scalable and reliable service for storing large amounts of data for longer duration
- Very simple web service interface for retrieving the data
- Once the CI logs are processed we can storage the logs for future reference in the S3 buckets with minimal cost.
- Access controls can be introduced in the granular level for more security.
- Buckets can be accessed both from console UI and programmatically
- In code we can RESTful architecture for accessing the buckets and objects in it.
- All the data is encrypted at rest for security.
- S3 buckets can be monitored by cloudTrail and cloudwatch

Benefits of S3:-

- It's a highly reliable and available service for accessing data from anywhere with high performance.
- It is durable, scalable and highly flexible, and comes at low cost.
- We can store and retrieve any type of data in it.
- Data is stored in the form of objects and imposing security is easy. We can implement bucket polices for access control.

- It can be connected to kinesis firehose or redshift for storing and retrieving data in real time.
- It can be used to run big data analytics and high performing computing applications.
- It offers various storage classes based on storage requirements.
- S3 glacier and s3 glacier deep archive storage classes for long term storage.
- It uses data encryption at rest and on the go for security and utilize versioning to restore previous data back.

Amazon redshift:-

- Data received form the amazon firehose will be fed to the redshift for real-time processing.
- Which in turn can be made available for the monitoring and visualization tools for analysis
- Historical data can be queried form the S3 data lake for processing without loading data.
- Query processing is fast by implanting techniques like columnar data storage and parallel processing with nodes.
- in multi node a leader node develops execution plans and runs on compute nodes for massive parallel processing

Features of Redshift:-

- It offers faster performance by employing various features like columnar data storage, advance compression techniques.
- As with multiple nodes deployed, it can perform parallel processing at massive scale and ideal for mission critical applications handling large chunk of data.
- It's very easy to setup, deploy and manage
- Data storage costs are very less and charged on hourly based
- It can be scaled up quickly as per requirement just by adding more nodes
- It's easily integrated with S3 Data Lake for querying date without loading using redshift spectrum.

- Redshift clusters can be connected using standard JDBC and ODBC drivers.
- It's highly secure with data encryption.

Amazon quick sight:-

- It's a business analytics tool provided by AWS to build visualizations and quickly analyze the data.
- It can include any sort of data either a third-party or inside AWS services.
- As it's a fully managed cloud based service it can provide enterprise grade security
- QuickSight used SPICE(Super-fast, Parallel, In-memory Calculation Engine) model for storing the prepared data
- Once the data is ready in redshift the same can be queried and visualized in the QuickSight
- And we can have multiple dashboards for current builds and the build history
- QuickSight provides dashboards for data visualization and various graph formats like column graph, line graphs, pie chart etc.
- Dashboards will be shared with multiple users at a time and which can be accessed using web UI.

Features of QuickSight:-

- It's a scalable , server less and machine learning powered business intelligence service
- It can handle multiple types of data either from third party or from native cloud services.
- It's a serve less and can automatically scale to tens and thousands of users.
- Its priced as pay per session based.
- It offers various data visualization tools and charts for preparing dashboards.
- Machine learning can be used to predict outcomes before head.
- Amazon QuickSight Q can be used to get answers for question using natural language.

- Reports can be shared in various forms like interactive dashboards and email reports and AWS console.

We can use various data warehouse OneSource tools as alternative for amazon redshift like mangoDB and PostgreSQL and big query from google and snowflake from snowflake Inc. can be used. For monitoring tools alternative we can use Nagios which is an open source and on premise tool which can be configured easily and scaled based on the infrastructure requirements. It's having rich tools for data monitoring and visualizations.