Big Data Analytics Options on AWS

AWS Whitepaper

Big Data Analytics Options on AWS: AWS Whitepaper

Table of Contents

Abstract 1
Introduction 2
The AWS Advantage in Big Data Analytics 3
Amazon Kinesis 4
Ideal Usage Patterns 5
Cost Model 5
Performance 5
Durability and Availability 6
Scalability and Elasticity 6
Interfaces 6
Anti-Patterns 6
AWS Lambda 6
Ideal Usage Patterns 7
Cost Model 7
Performance 7
Durability and Availability 7
Scalability and Elasticity 8
Interfaces 8
Anti-Patterns 8
Amazon EMR 8
Ideal Usage Patterns 9
Cost Model 9
Performance 9
Durability and Availability 9
Scalability and Elasticity 10
Interfaces 10
Anti-Patterns 12
AWS Glue 12
Ideal Usage Patterns 12
Cost Model 13
Performance 13
Durability and Availability 13
Scalability and Elasticity 13
Anti-Patterns 13
Amazon Machine Learning 14
Ideal Usage Patterns 14
Cost Model 14
Performance 15
Durability and Availability 15
Scalability and Elasticity 15
Interfaces 15
Anti-Patterns 16
Amazon DynamoDB 16
Ideal Usage Patterns 16
Cost Model 17
Performance 17
Durability and Availability 17
Scalability and Elasticity 17
Interfaces 18
Anti-Patterns 18
Amazon Redshift 18
Ideal Usage Patterns 19
Cost Model 19
Performance 19

Durability and Availability 20
Scalability and Elasticity 20
Interfaces 20
Anti-Patterns 20
Amazon Elasticsearch Service 21
Ideal Usage Patterns 21
Cost Model 22
Performance 22
Durability and Availability 22
Scalability and Elasticity 22
Interfaces 23
Anti-Patterns 23
Amazon QuickSight 23
Ideal Usage Patterns 24
Cost Model 24
Performance 24
Durability and Availability 25
Scalability and Elasticity 25
Interfaces 25
Anti-Patterns 25
Amazon EC2 25
Ideal Usage Patterns 26
Cost Model 26
Performance 26
Durability and Availability 26
Scalability and Elasticity 26
Interfaces 27
Anti-Patterns 27
Amazon Athena 27
Ideal Usage Patterns 27
Cost Model 28
Performance 28
Durability and Availability 28
Scalability and Elasticity 28
Security, Authorization and Encryption 28
Interfaces 29
Anti-Patterns 29
Solving Big Data Problems on AWS 30
Example 1: Queries against an Amazon S3 Data Lake 31
Example 2: Capturing and Analyzing Sensor Data 32
Example 3: Sentiment Analysis of Social Media 34
Conclusion 36
Contributors 37
Further Reading 38
Document Revisions 39
Notices 40

Big Data Analytics Options on AWS

Publication date: **December 2018** (*Document Revisions* (p. 39))

This whitepaper helps architects, data scientists, and developers understand the big data analytics options available in the AWS cloud by providing an overview of services, with the following information:

- Ideal usage patterns
- Cost model
- Performance
- Durability and availability
- Scalability and elasticity
- Interfaces
- Anti-patterns

This paper concludes with scenarios that showcase the analytics options in use, as well as additional resources for getting started with big data analytics on AWS.

Introduction

As we become a more digital society, the amount of data being created and collected is growing and accelerating significantly. Analysis of this ever-growing data becomes a challenge with traditional analytical tools. We require innovation to bridge the gap between data being generated and data that can be analyzed effectively.

Big data tools and technologies offer opportunities and challenges in being able to analyze data efficiently to better understand customer preferences, gain a competitive advantage in the marketplace, and grow your business. Data management architectures have evolved from the traditional data warehousing model to more complex architectures that address more requirements, such as real-time and batch processing; structured and unstructured data; high-velocity transactions; and so on.

Amazon Web Services (AWS) provides a broad platform of managed services to help you build, secure, and seamlessly scale end-to-end big data applications quickly and with ease. Whether your applications require real-time streaming or batch data processing, AWS provides the infrastructure and tools to tackle your next big data project. No hardware to procure, no infrastructure to maintain and scale—only what you need to collect, store, process, and analyze big data. AWS has an ecosystem of analytical solutions specifically designed to handle this growing amount of data and provide insight into your business.

The AWS Advantage in Big Data Analytics

Analyzing large data sets requires significant compute capacity that can vary in size based on the amount of input data and the type of analysis. This characteristic of big data workloads is ideally suited to the pay-as-you-go cloud computing model, where applications can easily scale up and down based on demand. As requirements change, you can easily resize your environment (horizontally or vertically) on AWS to meet your needs, without having to wait for additional hardware or being required to over invest to provision enough capacity.

For mission-critical applications on a more traditional infrastructure, system designers have no choice but to over-provision, because a surge in additional data due to an increase in business need must be something the system can handle. By contrast, on AWS you can provision more capacity and compute in a matter of minutes, meaning that your big data applications grow and shrink as demand dictates, and your system runs as close to optimal efficiency as possible.

In addition, you get flexible computing on a global infrastructure with access to the many different geographic regions that AWS offers, along with the ability to use other scalable services that augment to build sophisticated big data applications. These other services include Amazon Simple Storage Service (Amazon S3) to store data and AWS Glue to orchestrate jobs to move and transform that data easily. AWS IoT, which lets connected devices interact with cloud applications and other connected devices.

As the amount of data being generated continues to grow, AWS has many options to get that data to the cloud, including secure devices like AWS Snowball to accelerate petabyte-scale data transfers, delivery streams with Amazon Kinesis Data Firehose to load streaming data continuously, migrating databases using AWS Database Migration Service, and scalable private connections through AWS Direct Connect.

AWS recently added AWS Snowball Edge, which is a 100 TB data transfer device with on-board storage and compute capabilities. You can use Snowball Edge to move large amounts of data into and out of AWS, as a temporary storage tier for large local datasets, or to support local workloads in remote or offline locations. Additionally, you can deploy AWS Lambda code on Snowball Edge to perform tasks such as analyzing data streams or processing data locally.

As mobile continues to rapidly grow in usage you can use the suite of services within the AWS Mobile Hub to collect and measure app usage and data or export that data to another service for further custom analysis.

These capabilities of the AWS platform make it an ideal fit for solving big data problems, and many customers have implemented successful big data analytics workloads on AWS. For more information about case studies, see Big Data Customer Success Stories.

The following services for collecting, processing, storing, and analyzing big data are described in order:

Topics

- Amazon Kinesis (p. 4)
- AWS Lambda (p. 6)
- Amazon EMR (p. 8)
- AWS Glue (p. 12)
- Amazon Machine Learning (p. 14)
- Amazon DynamoDB (p. 16)
- Amazon Redshift (p. 18)
- Amazon Elasticsearch Service (p. 21)
- Amazon QuickSight (p. 23)

- Amazon EC2 (p. 25)
- Amazon Athena (p. 27)

In addition to these services, Amazon EC2 instances are available for self-managed big data applications.

Amazon Kinesis

Amazon Kinesis is a platform for streaming data on AWS, making it easy to load and analyze streaming data, and also providing the ability for you to build custom streaming data applications for specialized needs. With Kinesis, you can ingest real-time data such as application logs, website clickstreams, IoT telemetry data, and more into your databases, data lakes, and data warehouses, or build your own real-time applications using this data. Amazon Kinesis enables you to process and analyze data as it arrives and respond in real-time instead of having to wait until all your data is collected before the processing can begin.

Currently there are 4 pieces of the Kinesis platform that can be utilized based on your use case:

- Amazon Kinesis Data Streams enables you to build custom applications that process or analyze streaming data.
- Amazon Kinesis Video Streams enables you to build custom applications that process or analyze streaming video.
- Amazon Kinesis Data Firehose enables you to deliver real-time streaming data to AWS destinations such as Amazon S3, Amazon Redshift, Amazon Kinesis Analytics, and Amazon Elasticsearch Service.
- Amazon Kinesis Data Analytics enables you to process and analyze streaming data with standard SQL.

Kinesis Data Streams and Kinesis Video Streams enable you to build custom applications that process or analyze streaming data in real time. Kinesis Data Streams can continuously capture and store terabytes of data per hour from hundreds of thousands of sources, such as website clickstreams, financial transactions, social media feeds, IT logs, and location-tracking events. Kinesis Video Streams can continuously capture video data from smartphones, security cameras, drones, satellites, dashcams, and other edge devices.

With the Amazon Kinesis Client Library (KCL), you can build Amazon Kinesis applications and use streaming data to power real-time dashboards, generate alerts, and implement dynamic pricing and advertising. You can also emit data from Kinesis Data Streams and Kinesis Video Streams to other AWS services such as Amazon Simple Storage Service (Amazon S3), Amazon Redshift, Amazon Elastic MapReduce (Amazon EMR), and AWS Lambda.

Provision the level of input and output required for your data stream, in blocks of 1 megabyte per second (MB/sec), using the AWS Management Console, API, or SDKs. The size of your stream can be adjusted up or down at any time without restarting the stream and without any impact on the data sources pushing data to the stream. Within seconds, data put into a stream is available for analysis.

With Kinesis Data Firehose, you do not need to write applications or manage resources. You configure your data producers to send data to Kinesis Firehose and it automatically delivers the data to the AWS destination that you specified. You can also configure Kinesis Data Firehose to transform your data before data delivery. It is a fully managed service that automatically scales to match the throughput of your data and requires no ongoing administration. It can also batch, compress, and encrypt the data before loading it, minimizing the amount of storage used at the destination and increasing security.

Amazon Kinesis Data Analytics is the easiest way to process and analyze real-time, streaming data. With Kinesis Data Analytics, you just use standard SQL to process your data streams, so you don't have to learn any new programming languages. Simply point Kinesis Data Analytics at an incoming data stream, write your SQL queries, and specify where you want to load the results. Kinesis Data Analytics takes

care of running your SQL queries continuously on data while it's in transit and sending the results to the destinations.

In the subsequent sections we will focus primarily on Amazon Kinesis Data Streams.

Topics

- Ideal Usage Patterns (p. 5)
- Cost Model (p. 5)
- Performance (p. 5)
- Durability and Availability (p. 6)
- Scalability and Elasticity (p. 6)
- Interfaces (p. 6)
- Anti-Patterns (p. 6)

Ideal Usage Patterns

Amazon Kinesis Data Steams is useful wherever there is a need to move data rapidly off producers (data sources) and continuously process it. That processing can be to transform the data before emitting into another data store, drive real-time metrics and analytics, or derive and aggregate multiple streams into more complex streams, or downstream processing. The following are typical scenarios for using Kinesis Data Streams for analytics.

- **Real-time data analytics** –Kinesis Data Streams enables real-time data analytics on streaming data, such as analyzing website clickstream data and customer engagement analytics.
- **Log and data feed intake and processing** – With Kinesis Data Streams, you can have producers push data directly into an Amazon Kinesis stream. For example, you can submit system and application logs to Kinesis Data Streams and access the stream for processing within seconds. This prevents the log data from being lost if the front-end or application server fails, and reduces local log storage on the source. Kinesis Data Streams provides accelerated data intake because you are not batching up the data on the servers before you submit it for intake.
- **Real-time metrics and reporting** – You can use data ingested into Kinesis Data Streams for extracting metrics and generating KPIs to power reports and dashboards at real-time speeds. This enables data-processing application logic to work on data as it is streaming in continuously, rather than wait for data batches to arrive.

Cost Model

Amazon Kinesis Data Streams has simple pay-as-you-go pricing, with no up-front costs or minimum fees, and you only pay for the resources you consume. An Amazon Kinesis stream is made up of one or more shards, each shard gives you a capacity 5 read transactions per second, up to a maximum total of 2 MB of data read per second. Each shard can support up to 1,000 write transactions per second and up to a maximum total of 1 MB data written per second.

The data capacity of your stream is a function of the number of shards that you specify for the stream. The total capacity of the stream is the sum of the capacity of each shard. There are just two pricing components, an hourly charge per shard and a charge for each 1 million PUT transactions. For more information, see Amazon Kinesis Data Streams Pricing. Applications that run on Amazon EC2 and process Amazon Kinesis streams also incur standard Amazon EC2 costs.

Performance

Amazon Kinesis Data Streams allows you to choose throughput capacity you require in terms of shards. With each shard in an Amazon Kinesis stream, you can capture up to 1 megabyte per second of data at

1,000 write transactions per second. Your Amazon Kinesis applications can read data from each shard at up to 2 megabytes per second. You can provision as many shards as you need to get the throughput capacity you want; for instance, a 1 gigabyte per second data stream would require 1024 shards.

Durability and Availability

Amazon Kinesis Data Streams synchronously replicates data across three Availability Zones in an AWS Region, providing high availability and data durability.

Additionally, you can store a cursor in DynamoDB to durably track what has been read from an Amazon Kinesis stream. In the event that your application fails in the middle of reading data from the stream, you can restart your application and use the cursor to pick up from the exact spot where the failed application left off.

Scalability and Elasticity

You can increase or decrease the capacity of the stream at any time according to your business or operational needs, without any interruption to ongoing stream processing. By using API calls or development tools, you can automate scaling of your Amazon Kinesis Data Streams environment to meet demand and ensure you only pay for what you need.

Interfaces

There are two interfaces to Kinesis Data Streams: input which is used by data producers to put data into Kinesis Data Streams; and output to process and analyze data that comes in. Producers can write data using the Amazon Kinesis PUT API, an AWS Software Development Kit (SDK) or toolkit abstraction, the Amazon Kinesis Producer Library (KPL), or the Amazon Kinesis Agent.

For processing data that has already been put into an Amazon Kinesis stream, there are client libraries provided to build and operate real-time streaming data processing applications. The KCL17 acts as an intermediary between Amazon Kinesis Data Streams and your business applications which contain the specific processing logic. There is also integration to read from an Amazon Kinesis stream into Apache Storm via the Amazon Kinesis Storm Spout.

Anti-Patterns

Amazon Kinesis Data Streams has the following anti-patterns:

- **Small scale consistent throughput** – Even though Kinesis Data Streams works for streaming data at 200 KB/sec or less, it is designed and optimized for larger data throughputs.
- **Long-term data storage and analytics** –Kinesis Data Streams is not suited for long-term data storage. By default, data is retained for 24 hours, and you can extend the retention period by up to 7 days. You can move any data that needs to be stored for longer than 7 days into another durable storage service such as Amazon S3, Amazon Glacier, Amazon Redshift, or DynamoDB.

AWS Lambda

AWS Lambda

AWS Lambda lets you run code without provisioning or managing servers. You pay only for the compute time you consume – there is no charge when your code is not running. With Lambda, you can run code for virtually any type of application or backend service – all with zero administration. Just upload your code and Lambda takes care of everything required to run and scale your code with high availability. You can set up your code to automatically trigger from other AWS services or call it directly from any web or mobile app.

Topics

- Ideal Usage Patterns (p. 7)
- Cost Model (p. 7)
- Performance (p. 7)
- Durability and Availability (p. 7)
- Scalability and Elasticity (p. 8)
- Interfaces (p. 8)
- Anti-Patterns (p. 8)

Ideal Usage Patterns

AWS Lambda enables you to execute code in response to triggers such as changes in data, shifts in system state, or actions by users. Lambda can be directly triggered by AWS services such as Amazon S3, DynamoDB, Amazon Kinesis Data Streams, Amazon Simple Notification Service (Amazon SNS), and CloudWatch allowing you to build a variety of real-time data processing systems.

- **Real-time File Processing –** You can trigger Lambda to invoke a process where a file has been uploaded to Amazon S3 or modified. For example, to change an image from color to gray scale after it has been uploaded to Amazon S3.
- **Real-time Stream Processing** – You can use Kinesis Data Streams and Lambda to process streaming data for click stream analysis, log filtering, and social media analysis.
- **Extract, Transform, Load** – You can use Lambda to run code that transforms data and loads that data into one data repository to another.
- **Replace Cron** – Use schedule expressions to run a Lambda function at regular intervals as a cheaper and more available solution than running cron on an EC2 instance.
- **Process AWS Events** – Many other services, such as AWS CloudTrail, can act as event sources simply by logging to Amazon S3 and using S3 bucket notifications to trigger Lambda functions.

Cost Model

With AWS Lambda you only pay for what you use. You are charged based on the number of requests for your functions and the time your code executes. The Lambda free tier includes 1M free requests per month and 400,000 GB-seconds of compute time per month. You are charged $0.20 per 1 million requests thereafter ($0.0000002 per request). Additionally, the duration of your code executing is priced in relation to memory allocated. You are charged $0.00001667 for every GB-second used. See Lambda pricing for more details.

Performance

After deploying your code into Lambda for the first time, your functions are typically ready to call within seconds of upload. Lambda is designed to process events within milliseconds. Latency will be higher immediately after a Lambda function is created, updated, or if it has not been used recently. To improve performance, Lambda may choose to retain an instance of your function and reuse it to serve a subsequent request, rather than creating a new copy. To learn more about how Lambda reuses function instances, see our documentation. Your code should not assume that this reuse will always happen.

Durability and Availability

AWS Lambda is designed to use replication and redundancy to provide high availability for both the service itself and for the Lambda functions it operates. There are no maintenance windows or scheduled downtimes for either. On failure, Lambda functions being invoked synchronously respond with an

exception. Lambda functions being invoked asynchronously are retried at least 3 times, after which the event may be rejected.

Scalability and Elasticity

There is no limit on the number of Lambda functions that you can run. However, Lambda has a default safety throttle of 1,000 concurrent executions per account per region. A member of the AWS support team can increase this limit. Lambda is designed to scale automatically on your behalf. There are no fundamental limits to scaling a function. Lambda dynamically allocates capacity to match the rate of incoming events.

Interfaces

Lambda functions can be managed in a variety of ways. You can easily list, delete, update, and monitor your Lambda functions using the dashboard in the Lambda console. You also can use the AWS CLI and AWS SDK to manage your Lambda functions.

You can trigger a Lambda function from an AWS event, such as Amazon S3 bucket notifications, Amazon DynamoDB Streams, Amazon CloudWatch logs, Amazon Simple Email Service (Amazon SES), Amazon Kinesis Data Streams, Amazon SNS, Amazon Cognito, and more. Any API call in any service that supports AWS CloudTrail can be processed as an event in Lambda by responding to CloudTrail audit logs. For more information about event sources, see Core Components: AWS Lambda Function and Event Sources.

AWS Lambda supports code written in Node.js (JavaScript), Python, Java (Java 8 compatible), C# (.NET Core), Go, PowerShell and Ruby. Your code can include existing libraries, even native ones. Please read our documentation on using Node.js, Python, Java, C#, Go, PowerShell and Ruby.

Anti-Patterns

- **Long Running Applications –** Each Lambda function must complete within 900 seconds. For long running applications that may require jobs to run longer than fifteen minutes, Amazon EC2 is recommended. Alternately, create a chain of Lambda functions where function 1, calls function 2, which calls function 3, and so on, until the process is completed. See Creating a Lambda State Machine for more information.
- **Dynamic Websites –** While it is possible to run a static website with AWS Lambda, running a highly dynamic and large volume website can be performance prohibitive. Utilizing Amazon EC2 and Amazon CloudFront would be a recommended use-case.
- **Stateful Applications –**Lambda code must be written in a “stateless” style, i.e., it should assume there is no affinity to the underlying compute infrastructure. Local file system access, child processes, and similar artifacts may not extend beyond the lifetime of the request, and any persistent state should be stored in Amazon S3, DynamoDB, or another Internet-available storage service.

Amazon EMR

Amazon EMR is a highly distributed computing framework to easily process and store data quickly in a cost-effective manner. Amazon EMR uses Apache Hadoop, an open source framework, to distribute your data and processing across a resizable cluster of Amazon EC2 instances and allows you to use the most common Hadoop tools such as Hive, Pig, Spark, and so on. Hadoop provides a framework to run big data processing and analytics. Amazon EMR does all the work involved with provisioning, managing, and maintaining the infrastructure and software of a Hadoop cluster.

Topics

- Ideal Usage Patterns (p. 9)

- Cost Model (p. 9)
- Performance (p. 9)
- Durability and Availability (p. 9)
- Scalability and Elasticity (p. 10)
- Interfaces (p. 10)
- Anti-Patterns (p. 12)

Ideal Usage Patterns

Amazon EMR's flexible framework reduces large processing problems and data sets into smaller jobs and distributes them across many compute nodes in a Hadoop cluster. This capability lends itself to many usage patterns with big data analytics. Here are a few examples:

- Log processing and analytics
- Large extract, transform, and load (ETL) data movement
- Risk modeling and threat analytics
- Ad targeting and click stream analytics
- Genomics
- Predictive analytics
- Ad hoc data mining and analytics

For more information, see the documentation for Amazon EMR.

Cost Model

With Amazon EMR, you can launch a persistent cluster that stays up indefinitely or a temporary cluster that terminates after the analysis is complete. In either scenario, you only pay for the hours the cluster is up.

Amazon EMR supports a variety of Amazon EC2 instance types (standard, high CPU, high memory, high I/O, and so on) and all Amazon EC2 pricing options (On-Demand, Reserved, and Spot). When you launch an Amazon EMR cluster (also called a "job flow"), you choose how many and what type of Amazon EC2 instances to provision. The Amazon EMR price is in addition to the Amazon EC2 price. For more information, see Amazon EMR Pricing.

Performance

Amazon EMR performance is driven by the type of EC2 instances you choose to run your cluster on and how many you chose to run your analytics. You should choose an instance type suitable for your processing requirements, with sufficient memory, storage, and processing power. For more information about EC2 instance specifications, see Amazon EC2 Instance Types.

Durability and Availability

By default, Amazon EMR is fault tolerant for core node failures and continues job execution if a slave node goes down. Amazon EMR will also provision a new node when a core node fails. However, Amazon EMR will not replace nodes if all nodes in the cluster are lost. Customers can monitor the health of nodes and replace failed nodes with CloudWatch.

Amazon EMR is fault tolerant for slave failures and continues job execution if a slave node goes down. Amazon EMR will also provision a new node when a core node fails. However, Amazon EMR will not replace nodes if all nodes in the cluster are lost.

Scalability and Elasticity

With Amazon EMR, it is easy to resize a running cluster. You can add core nodes which hold the Hadoop Distributed File System (HDFS) at any time to increase your processing power and increase the HDFS storage capacity (and throughput). Additionally, you can use Amazon S3 natively or using EMRFS along with or instead of local HDFS which allows you to decouple your memory and compute from your storage providing greater flexibility and cost efficiency.

You can also add and remove task nodes at any time which can process Hadoop jobs, but do not maintain HDFS. Some customers add hundreds of instances to their clusters when their batch processing occurs, and remove the extra instances when processing completes. For example, you may not know how much data your clusters will be handling in 6 months, or you may have spikey processing needs.

With Amazon EMR, you don't need to guess your future requirements or provision for peak demand because you can easily add or remove capacity at any time.

Additionally, you can add all new clusters of various sizes and remove them at any time with a few clicks in the console or by a programmatic API call.

Interfaces

Amazon EMR supports many tools on top of Hadoop that can be used for big data analytics and each has their own interfaces. Here is a brief summary of the most popular options:

Hive

Hive is an open source data warehouse and analytics package that runs on top of Hadoop. Hive is operated by Hive QL, a SQL-based language which allows users to structure, summarize, and query data. Hive QL goes beyond standard SQL, adding first-class support for map/reduce functions and complex extensible user-defined data types like JSON and Thrift. This capability allows processing of complex and unstructured data sources such as text documents and log files.

Hive allows user extensions via user-defined functions written in Java. Amazon EMR has made numerous improvements to Hive, including direct integration with DynamoDB and Amazon S3. For example, with Amazon EMR you can load table partitions automatically from Amazon S3, you can write data to tables in Amazon S3 without using temporary files, and you can access resources in Amazon S3, such as scripts for custom map and/or reduce operations and additional libraries. For more information, see Apache Hive in the *Amazon EMR Release Guide*.

Pig

Pig is an open source analytics package that runs on top of Hadoop. Pig is operated by Pig Latin, a SQL-like language which allows users to structure, summarize, and query data. As well as SQL-like operations, Pig Latin also adds first-class support for map and reduce functions and complex extensible user defined data types. This capability allows processing of complex and unstructured data sources such as text documents and log files.

Pig allows user extensions via user-defined functions written in Java. Amazon EMR has made numerous improvements to Pig, including the ability to use multiple file systems (normally, Pig can only access one remote file system), the ability to load customer JARs and scripts from Amazon S3 (such as “REGISTER s3://my-bucket/piggybank.jar”), and additional functionality for String and DateTime processing. For more information, see Apache Pig33 in the *Amazon EMR Release Guide*.

Spark

Spark is an open source data analytics engine built on Hadoop with the fundamentals for in-memory MapReduce. Spark provides additional speed for certain analytics and is the foundation for other power

tools such as Shark (SQL driven data warehousing), Spark Streaming (streaming applications), GraphX (graph systems) and MLlib (machine learning). For more information, see Apache Spark on Amazon EMR .

HBase

HBase is an open source, non-relational, distributed database modeled after Google's BigTable. It was developed as part of Apache Software Foundation's Hadoop project and runs on top of Hadoop Distributed File System (HDFS) to provide BigTable-like capabilities for Hadoop. HBase provides you a fault-tolerant, efficient way of storing large quantities of sparse data using column-based compression and storage. In addition, HBase provides fast lookup of data because data is stored in-memory instead of on disk.

HBase is optimized for sequential write operations, and it is highly efficient for batch inserts, updates, and deletes. HBase works seamlessly with Hadoop, sharing its file system and serving as a direct input and output to Hadoop jobs. HBase also integrates with Apache Hive, enabling SQL-like queries over HBase tables, joins with Hive-based tables, and support for Java Database Connectivity (JDBC). With Amazon EMR, you can back up HBase to Amazon S3 (full or incremental, manual or automated) and you can restore from a previously created backup. For more information, see Apache HBase in the *Amazon EMR Release Guide*.

Hunk

Hunk was developed by Splunk to make machine data accessible, usable, and valuable to everyone. With Hunk, you can interactively explore, analyze, and visualize data stored in Amazon EMR and Amazon S3, harnessing Splunk analytics on Hadoop. For more information, see Amazon EMR with Hunk: Splunk Analytics for Hadoop and NoSQL.

Presto

Presto is an open-source distributed SQL query engine optimized for low-latency, ad-hoc analysis of data. It supports the ANSI SQL standard, including complex queries, aggregations, joins, and window functions. Presto can process data from multiple data sources including the Hadoop Distributed File System (HDFS) and Amazon S3.

Kinesis Connector

The Kinesis Connector enables EMR to directly read and query data from Kinesis Data Streams. You can perform batch processing of Kinesis streams using existing Hadoop ecosystem tools such as Hive, Pig, MapReduce, Hadoop Streaming, and Cascading. Some use cases enabled by this integration are:

- Streaming Log Analysis: You can analyze streaming web logs to generate a list of top 10 error type every few minutes by region, browser, and access domains.
- Complex Data Processing Workflows: You can join Kinesis stream with data stored in Amazon S3, Dynamo DB tables, and HDFS. You can write queries that join clickstream data from Kinesis with advertising campaign information stored in a DynamoDB table to identify the most effective categories of ads that are displayed on particular websites.
- Ad-hoc Queries: You can periodically load data from Kinesis into HDFS and make it available as a local Impala table for fast, interactive, analytic queries.

Other third-party tools

Amazon EMR also supports a variety of other popular applications and tools in the Hadoop ecosystem, such as R (statistics), Mahout (machine learning), Ganglia (monitoring), Accumulo (secure NoSQL database), Hue (user interface to analyze Hadoop data), Sqoop (relational database connector), HCatalog (table and storage management), and more.

Additionally, you can install your own software on top of Amazon EMR to help solve your business needs. AWS provides the ability to quickly move large amounts of data from Amazon S3 to HDFS, from HDFS to Amazon S3, and between Amazon S3 buckets using Amazon EMR's S3DistCp, an extension of the open source tool DistCp that uses MapReduce to efficiently move large amounts of data.

You can optionally use the EMR File System (EMRFS), an implementation of HDFS which allows Amazon EMR clusters to store data on Amazon S3. You can enable Amazon S3 server-side and client-side encryption. When you use EMRFS, a metadata store is transparently built in DynamoDB to help manage the interactions with Amazon S3 and allows you to have multiple EMR clusters easily use the same EMRFS metadata and storage on Amazon S3.

Anti-Patterns

Amazon EMR has the following anti-patterns:

- **Small data sets** – Amazon EMR is built for massive parallel processing; if your data set is small enough to run quickly on a single machine, in a single thread, the added overhead to map and reduce jobs may not be worth it for small data sets that can easily be processed in memory on a single system.
- **ACID transaction requirements** – While there are ways to achieve ACID (atomicity, consistency, isolation, durability) or limited ACID on Hadoop, using another database, such as Amazon Relational Database Service (Amazon RDS) or a relational database running on Amazon EC2 may be a better option for workloads with stringent requirements.

AWS Glue

AWS Glue is a fully managed extract, transform, and load (ETL) service that you can use to catalog your data, clean it, enrich it, and move it reliably between data stores. With AWS Glue, you can significantly reduce the cost, complexity, and time spent creating ETL jobs. AWS Glue is Serverless, so there is no infrastructure to setup or manage. You pay only for the resources consumed while your jobs are running.

Topics

- Ideal Usage Patterns (p. 12)
- Cost Model (p. 13)
- Performance (p. 13)
- Durability and Availability (p. 13)
- Scalability and Elasticity (p. 13)
- Anti-Patterns (p. 13)

Ideal Usage Patterns

AWS Glue is designed to easily prepare data for extract, transform, and load (ETL) jobs. Using AWS Glue gives you the following benefits:

- AWS Glue can automatically crawl your data and generate code to execute or data transformations and loading processes.
- Integration with services like Amazon Athena, Amazon EMR, and Amazon Redshift
- Serverless, no infrastructure to provision or manage
- AWS Glue generates ETL code that is customizable, reusable, and portable, using familiar technology – Python and Spark.

Cost Model

With AWS Glue, you pay an hourly rate, billed by the minute, for crawler jobs (discovering data) and ETL jobs (processing and loading data). For the AWS 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. If you provision a development endpoint to interactively develop your ETL code, you pay an hourly rate, billed per minute. See AWS Glue Pricing for more details.

Performance

AWS Glue uses a scale-out Apache Spark environment to load your data into its destination. You can simply specify the number of Data Processing Units (DPUs) that you want to allocate to your ETL job. An AWS Glue ETL job requires a minimum of 2 DPUs. By default, AWS Glue allocates 10 DPUs to each ETL job. Additional DPUs can be added to increase the performance of your ETL job. Multiple jobs can be triggered in parallel or sequentially by triggering them on a job completion event. You can also trigger one or more AWS Glue jobs from an external source such as an AWS Lambda function.

Durability and Availability

AWS Glue connects to the data source of your preference, whether it is in an Amazon S3 file, an Amazon RDS table, or another set of data. As a result, all your data is stored and available as it pertains to that data stores durability characteristics. The AWS Glue service provides status of each job and pushes all notifications to Amazon CloudWatch events. You can setup SNS notifications using CloudWatch actions to be informed of job failures or completions.

Scalability and Elasticity

AWS Glue provides a managed ETL service that runs on a Serverless Apache Spark environment. This allows you to focus on your ETL job and not worry about configuring and managing the underlying compute resources. AWS Glue works on top of the Apache Spark environment to provide a scale-out execution environment for your data transformation jobs.

Interfaces

AWS Glue provides a number of ways to populate metadata into the AWS AWS Glue Data Catalog. AWS Glue crawlers scan various data stores you own to automatically infer schemas and partition structure and populate the AWS AWS Glue Data Catalog with corresponding table definitions and statistics. You can also schedule crawlers to run periodically so that your metadata is always up-to-date and in-sync with the underlying data. Alternately, you can add and update table details manually by using the AWS Glue Console or by calling the API. You can also run Hive DDL statements via the Amazon Athena Console or a Hive client on an Amazon EMR cluster. Finally, if you already have a persistent Apache Hive Metastore, you can perform a bulk import of that metadata into the AWS AWS Glue Data Catalog by using our import script.

Anti-Patterns

AWS Glue has the following anti-patterns:

- **Streaming data** – AWS Glue ETL is batch oriented, and you can schedule your ETL jobs at a minimum of 5 minute intervals. While it can process micro-batches, it does not handle streaming data. If your use case requires you to ETL data while you stream it in, you can perform the first leg of your ETL using Amazon Kinesis, Amazon Kinesis Data Firehose, or Amazon Kinesis Analytics. Then store the data in either Amazon S3 or Amazon Redshift and trigger an AWS Glue ETL job to pick up that dataset and continue applying additional transformations to that data.

- **Multiple ETL engines** – AWS Glue ETL jobs are PySpark based. If your use case requires you to use an engine other than Apache Spark or if you want to run a heterogeneous set of jobs that run on a variety of engines like Hive, Pig, etc., then AWS Data Pipeline or Amazon EMR would be a better choice.
- **NoSQL Databases** – Currently AWS Glue does not support data sources like NoSQL databases or Amazon DynamoDB. Since NoSQL databases do not require a rigid schema like traditional relational databases, most common ETL jobs would not apply.

Amazon Machine Learning

Amazon Machine Learning (Amazon ML) is a service that makes it easy for anyone to use predictive analytics and machine-learning technology. Amazon ML provides visualization tools and wizards to guide you through the process of creating machine learning (ML) models without having to learn complex ML algorithms and technology. After your models are ready, Amazon ML makes it easy to obtain predictions for your application using API operations, without having to implement custom prediction generation code or manage any infrastructure.

Amazon ML can create ML models based on data stored in Amazon S3, Amazon Redshift, or Amazon RDS. Built-in wizards guide you through the steps of interactively exploring your data, to training the ML model, to evaluating the model quality and adjusting outputs to align with business goals. After a model is ready, you can request predictions in either batches or using the low-latency real-time API.

Topics

- Ideal Usage Patterns (p. 14)
- Cost Model (p. 14)
- Performance (p. 15)
- Durability and Availability (p. 15)
- Scalability and Elasticity (p. 15)
- Interfaces (p. 15)
- Anti-Patterns (p. 16)

Ideal Usage Patterns

Amazon ML is ideal for discovering patterns in your data and using these patterns to create ML models that can generate predictions on new, unseen data points. For example, you can:

- **Enable applications to flag suspicious transactions** – Build an ML model that predicts whether a new transaction is legitimate or fraudulent.
- **Forecast product demand** – Input historical order information to predict future order quantities.
- **Personalize application content** – Predict which items a user will be most interested in, and retrieve these predictions from your application in real-time.
- **Predict user activity** – Analyze user behavior to customize your website and provide a better user experience.
- **Listen to social media** – Ingest and analyze social media feeds that potentially impact business decisions.

Cost Model

With Amazon ML, you pay only for what you use. There are no minimum fees and no upfront commitments. Amazon ML charges an hourly rate for the compute time used to build predictive models, and then you pay for the number of predictions generated for your application. For real-time predictions

you also pay an hourly reserved capacity charge based on the amount of memory required to run your model.

The charge for data analysis, model training, and evaluation is based on the number of compute hours required to perform them, and depends on the size of the input data, the number of attributes within it, and the number and types of transformations applied. Data analysis and model building fees are priced at $0.42 per hour. Prediction fees are categorized as batch and real-time. Batch predictions are $0.10 per 1,000 predictions, rounded up to the next 1,000, while real-time predictions are $0.0001 per prediction, rounded up to the nearest penny. For real-time predictions, there is also a reserved capacity charge of $0.001 per hour for each 10 MB of memory provisioned for your model.

During model creation, you specify the maximum memory size of each model to manage cost and control predictive performance. You pay the reserved capacity charge only while your model is enabled for real-time predictions. Charges for data stored in Amazon S3, Amazon RDS, or Amazon Redshift are billed separately. For more information, see Amazon Machine Learning Pricing.

Performance

The time it takes to create models, or to request batch predictions from these models depends on the number of input data records, the types and distribution of attributes within these records, and the complexity of the data processing "recipe" that you specify.

Most real-time prediction requests return a response within 100 ms, making them fast enough for interactive web, mobile, or desktop applications. The exact time it takes for the real-time API to generate a prediction varies depending on the size of the input data record, and the complexity of the data processing "recipe" associated with the ML model that is generating the predictions. Each ML model that is enabled for real-time predictions can be used to request up to 200 transactions per second by default, and this number can be increased by contacting customer support. You can monitor the number of predictions requested by your ML models by using CloudWatch metrics.

Durability and Availability

Amazon ML is designed for high availability. There are no maintenance windows or scheduled downtimes. The service runs in Amazon's proven, high-availability data centers, with service stack replication configured across three facilities in each AWS Region to provide fault tolerance in the event of a server failure or Availability Zone outage.

Scalability and Elasticity

By default, you can process data sets that are up to 100 GB (this can be increased with a support ticket) in size to create ML models or to request batch predictions. For large volumes of batch predictions, you can split your input data records into separate chunks to enable the processing of larger prediction data volume.

By default, you can run up to five simultaneous jobs and by contacting customer service you can have this limit raised. Because Amazon ML is a managed service, there are no servers to provision and as a result you are able to scale as your application grows without having to over provision or pay for resources not being used.

Interfaces

Creating a data source is as simple as adding your data to Amazon S3 or you can pull data directly from Amazon Redshift or MySQL databases managed by Amazon RDS. After your data source is defined, you can interact with Amazon ML using the console. Programmatic access to Amazon ML is enabled by the AWS SDKs and Amazon ML API. You can also create and manage Amazon ML entities using the AWS CLI available on Windows, Mac, and Linux/UNIX systems.

Anti-Patterns

Amazon ML has the following anti-patterns:

- **Very large data sets** – While Amazon ML can support up to a default 100 GB of data (this can be increased with a support ticket), terabyte-scale ingestion of data is not currently supported. Using Amazon EMR to run Spark's Machine Learning Library (MLlib) is a common tool for such a use case.
- **Unsupported learning tasks** – Amazon ML can be used to create ML models that perform binary classification (choose one of two choices, and provide a measure of confidence), multiclass classification (extend choices to beyond two options), or numeric regression (predict a number directly). Unsupported ML tasks such as sequence prediction or unsupervised clustering can be approached by using Amazon EMR to run Spark and MLlib.

Amazon DynamoDB

Amazon DynamoDB is a fast, fully-managed NoSQL database service that makes it simple and cost effective to store and retrieve any amount of data, and serve any level of request traffic. DynamoDB helps offload the administrative burden of operating and scaling a highly-available distributed database cluster. This storage alternative meets the latency and throughput requirements of highly demanding applications by providing single-digit millisecond latency and predictable performance with seamless throughput and storage scalability.

DynamoDB stores structured data in tables, indexed by primary key, and allows low-latency read and write access to items ranging from 1 byte up to 400 KB. DynamoDB supports three data types (number, string, and binary), in both scalar and multi-valued sets. It supports document stores such as JSON, XML, or HTML in these data types. Tables do not have a fixed schema, so each data item can have a different number of attributes. The primary key can either be a single-attribute hash key or a composite hash-range key.

DynamoDB offers both global and local secondary indexes provide additional flexibility for querying against attributes other than the primary key. DynamoDB provides both eventually-consistent reads (by default), and strongly-consistent reads (optional), as well as implicit item-level transactions for item put, update, delete, conditional operations, and increment/decrement.

DynamoDB is integrated with other services, such as Amazon EMR, Amazon Redshift, AWS Data Pipeline, and Amazon S3, for analytics, data warehouse, data import/export, backup, and archive.

Topics

- Ideal Usage Patterns (p. 16)
- Cost Model (p. 17)
- Performance (p. 17)
- Durability and Availability (p. 17)
- Scalability and Elasticity (p. 17)
- Interfaces (p. 18)
- Anti-Patterns (p. 18)

Ideal Usage Patterns

DynamoDB is ideal for existing or new applications that need a flexible NoSQL database with low read and write latencies, and the ability to scale storage and throughput up or down as needed without code changes or downtime.

Common use cases include:

- Mobile apps
- Gaming
- Digital ad serving
- Live voting
- Audience interaction for live events
- Sensor networks
- Log ingestion
- Access control for web-based content
- Metadata storage for Amazon S3 objects
- E-commerce shopping carts
- Web session management

Many of these use cases require a highly available and scalable database because downtime or performance degradation has an immediate negative impact on an organization's business.

Cost Model

With DynamoDB, you pay only for what you use and there is no minimum fee. DynamoDB has three pricing components: provisioned throughput capacity (per hour), indexed data storage (per GB per month), data transfer in or out (per GB per month). New customers can start using DynamoDB for free as part of the AWS Free Usage Tier. For more information, see Amazon DynamoDB Pricing.

Performance

SSDs and limiting indexing on attributes provides high throughput and low latency and drastically reduces the cost of read and write operations. As the datasets grow, predictable performance is required so that low-latency for the workloads can be maintained. This predictable performance can be achieved by defining the provisioned throughput capacity required for a given table.

Behind the scenes, the service handles the provisioning of resources to achieve the requested throughput rate; you don't need to think about instances, hardware, memory, and other factors that can affect an application's throughput rate. Provisioned throughput capacity reservations are elastic and can be increased or decreased on demand.

Durability and Availability

DynamoDB has built-in fault tolerance that automatically and synchronously replicates data across three data centers in a region for high availability and to help protect data against individual machine, or even facility, failures.

Amazon DynamoDB Streams captures all data activity that happens on your table and allows the ability to set up regional replication from one geographic region to another to provide even greater availability.

Scalability and Elasticity

DynamoDB is both highly scalable and elastic. There is no limit to the amount of data that you can store in a DynamoDB table, and the service automatically allocates more storage as you store more data using the DynamoDB write API operations. Data is automatically partitioned and re-partitioned as needed, while the use of SSDs provides predictable low-latency response times at any scale. The service is also

elastic, in that you can simply “dial-up” or “dial-down” the read and write capacity of a table as your needs change.

Interfaces

DynamoDB provides a low-level REST API, as well as higher-level SDKs for Java, ET, and PHP that wrap the low-level REST API and provide some object-relational mapping (ORM) functions. These APIs provide both a management and data interface for DynamoDB. The API currently offers operations that enable table management (creating, listing, deleting, and obtaining metadata) and working with attributes (getting, writing, and deleting attributes; query using an index, and full scan).

While standard SQL isn’t available, you can use the DynamoDB select operation to create SQL-like queries that retrieve a set of attributes based on criteria that you provide. You can also work with DynamoDB using the console.

Anti-Patterns

DynamoDB has the following anti-patterns:

- **Prewritten application tied to a traditional relational database** – If you are attempting to port an existing application to the AWS cloud and need to continue using a relational database, you can use either Amazon RDS (Amazon Aurora, MySQL, PostgreSQL, Oracle, or SQL Server), or one of the many pre-configured Amazon EC2 database AMIs. You can also install your choice of database software on an EC2 instance that you manage.
- **Joins or complex transactions** – While many solutions are able to leverage DynamoDB to support their users, it’s possible that your application may require joins, complex transactions, and other relational infrastructure provided by traditional database platforms. If this is the case, you may want to explore Amazon Redshift, Amazon RDS, or Amazon EC2 with a self-managed database.
- **Binary large objects (BLOB) data** – If you plan on storing large (greater than 400 KB) BLOB data, such as digital video, images, or music, you’ll want to consider Amazon S3. However, DynamoDB can be used in this scenario for keeping track of metadata (e.g., item name, size, date created, owner, location, etc.) about your binary objects.
- **Large data with low I/O rate** –DynamoDB uses SSD drives and is optimized for workloads with a high I/O rate per GB stored. If you plan to store very large amounts of data that are infrequently accessed, other storage options may be a better choice, such as Amazon S3.

Amazon Redshift

Amazon Redshift is a fast, fully-managed, petabyte-scale data warehouse service that makes it simple and cost-effective to analyze all your data efficiently using your existing business intelligence tools. It is optimized for data sets ranging from a few hundred gigabytes to a petabyte or more, and is designed to cost less than a tenth of the cost of most traditional data warehousing solutions.

Amazon Redshift delivers fast query and I/O performance for virtually any size dataset by using columnar storage technology while parallelizing and distributing queries across multiple nodes. It automates most of the common administrative tasks associated with provisioning, configuring, monitoring, backing up, and securing a data warehouse, making it easy and inexpensive to manage and maintain. This automation allows you to build petabyte-scale data warehouses in minutes instead of weeks or months taken by traditional on-premises implementations.

Amazon Redshift Spectrum is a feature that enables you to run queries against exabytes of unstructured data in Amazon S3, with no loading or ETL required. When you issue a query, it goes to the Amazon Redshift SQL endpoint, which generates and optimizes a query plan. Amazon Redshift determines what

data is local and what is in Amazon S3, generates a plan to minimize the amount of Amazon S3 data that needs to be read, and then requests Redshift Spectrum workers out of a shared resource pool to read and process the data from Amazon S3.

Topics

- Ideal Usage Patterns (p. 19)
- Cost Model (p. 19)
- Performance (p. 19)
- Durability and Availability (p. 20)
- Scalability and Elasticity (p. 20)
- Interfaces (p. 20)
- Anti-Patterns (p. 20)

Ideal Usage Patterns

Amazon Redshift is ideal for online analytical processing (OLAP) using your existing business intelligence tools. Organizations are using Amazon Redshift to:

- Analyze global sales data for multiple products
- Store historical stock trade data
- Analyze ad impressions and clicks
- Aggregate gaming data
- Analyze social trends
- Measure clinical quality, operation efficiency, and financial performance in health care

Cost Model

An Amazon Redshift data warehouse cluster requires no long-term commitments or upfront costs. This frees you from the capital expense and complexity of planning and purchasing data warehouse capacity ahead of your needs. Charges are based on the size and number of nodes of your cluster.

There is no additional charge for backup storage up to 100% of your provisioned storage. For example, if you have an active cluster with 2 XL nodes for a total of 4 TB of storage, AWS provides up to 4 TB of backup storage on Amazon S3 at no additional charge. Backup storage beyond the provisioned storage size, and backups stored after your cluster is terminated, are billed at standard Amazon S3 rates. There is no data transfer charge for communication between Amazon S3 and Amazon Redshift. For more information, see Amazon Redshift pricing.

Performance

Amazon Redshift uses a variety of innovations to obtain very high performance on data sets ranging in size from hundreds of gigabytes to a petabyte or more. It uses columnar storage, data compression, and zone maps to reduce the amount of I/O needed to perform queries.

Amazon Redshift has a massively parallel processing (MPP) architecture, parallelizing and distributing SQL operations to take advantage of all available resources. The underlying hardware is designed for high performance data processing, using local attached storage to maximize throughput between the CPUs and drives, and a 10 GigE mesh network to maximize throughput between nodes. Performance can be tuned based on your data warehousing needs: AWS offers Dense Compute (DC) with SSD drives as well as Dense Storage (DS) options.

Durability and Availability

Amazon Redshift automatically detects and replaces a failed node in your data warehouse cluster. The data warehouse cluster is read-only until a replacement node is provisioned and added to the DB, which typically only takes a few minutes. Amazon Redshift makes your replacement node available immediately and streams your most frequently accessed data from Amazon S3 first to allow you to resume querying your data as quickly as possible.

Additionally, your data warehouse cluster remains available in the event of a drive failure; because Amazon Redshift mirrors your data across the cluster, it uses the data from another node to rebuild failed drives. Amazon Redshift clusters reside within one Availability Zone, but if you wish to have a multi-AZ set up for Amazon Redshift, you can set up a mirror and then self-manage replication and failover.

Scalability and Elasticity

With a few clicks in the console or an API call, you can easily change the number, or type, of nodes in your data warehouse as your performance or capacity needs change. Amazon Redshift enables you to start with a single 160 GB node and scale up to a petabyte or more of compressed user data using many nodes. For more information, see Clusters and Nodes in Amazon Redshift in the *Amazon Redshift Management Guide*.

While resizing, Amazon Redshift places your existing cluster into read-only mode, provisions a new cluster of your chosen size, and then copies data from your old cluster to your new one in parallel. During this process, you pay only for the active Amazon Redshift cluster. You can continue running queries against your old cluster while the new one is being provisioned. After your data has been copied to your new cluster, Amazon Redshift automatically redirects queries to your new cluster and removes the old cluster.

Interfaces

Amazon Redshift has custom JDBC and ODBC drivers that you can download from the Connect Client tab of the console, allowing you to use a wide range of familiar SQL clients. You can also use standard PostgreSQL JDBC and ODBC drivers. For more information about Amazon Redshift drivers, see Amazon Redshift and PostgreSQL.

There are numerous examples of validated integrations with many popular BI and ETL vendors. Loads and unloads are attempted in parallel into each compute node to maximize the rate at which you can ingest data into your data warehouse cluster as well as to and from Amazon S3 and DynamoDB. You can easily load streaming data into Amazon Redshift using Amazon Kinesis Data Firehose, enabling near real-time analytics with existing business intelligence tools and dashboards you're already using today. Metrics for compute utilization, memory utilization, storage utilization, and read/write traffic to your Amazon Redshift data warehouse cluster are available free of charge via the console or CloudWatch API operations.

Anti-Patterns

Amazon Redshift has the following anti-patterns:

- **Small data sets** – Amazon Redshift is built for parallel processing across a cluster. If your data set is less than a hundred gigabytes, you are not going to get all the benefits that Amazon Redshift has to offer and Amazon RDS may be a better solution.
- **On-line transaction processing (OLTP)** – Amazon Redshift is designed for data warehouse workloads producing extremely fast and inexpensive analytic capabilities. If you require a fast transactional

system, you may want to choose a traditional relational database system built on Amazon RDS or a NoSQL database offering, such as DynamoDB.

- **Unstructured data** – Data in Amazon Redshift must be structured by a defined schema, rather than supporting arbitrary schema structure for each row. If your data is unstructured, you can perform extract, transform, and load (ETL) on Amazon EMR to get the data ready for loading into Amazon Redshift.
- **BLOB data** – If you plan on storing large binary files (such as digital video, images, or music), you may want to consider storing the data in Amazon S3 and referencing its location in Amazon Redshift. In this scenario, Amazon Redshift keeps track of metadata (such as item name, size, date created, owner, location, and so on) about your binary objects, but the large objects themselves are stored in Amazon S3.

Amazon Elasticsearch Service

Amazon Elasticsearch Service (Amazon ES) makes it easy to deploy, operate, and scale Elasticsearch for log analytics, full text search, application monitoring, and more. Amazon ES is a fully managed service that delivers Elasticsearch's easy-to-use APIs and real-time capabilities along with the availability, scalability, and security required by production workloads. The service offers built-in integrations with Kibana, Logstash, and AWS services including Amazon Kinesis Data Firehose, AWS Lambda, and Amazon CloudWatch so that you can go from raw data to actionable insights quickly.

It's easy to get started with Amazon ES. You can set up and configure your Amazon ES domain in minutes from the AWS Management Console. Amazon ES provisions all the resources for your domain and launches it. The service automatically detects and replaces failed Elasticsearch nodes, reducing the overhead associated with self-managed infrastructure and Elasticsearch software. Amazon ES allows you to easily scale your cluster via a single API call or a few clicks in the console. With Amazon ES, you get direct access to the Elasticsearch open-source API so that code and applications you're already using with your existing Elasticsearch environments will work seamlessly.

Topics

- Ideal Usage Patterns (p. 21)
- Cost Model (p. 22)
- Performance (p. 22)
- Durability and Availability (p. 22)
- Scalability and Elasticity (p. 22)
- Interfaces (p. 23)
- Anti-Patterns (p. 23)

Ideal Usage Patterns

Amazon Elasticsearch Service is ideal for querying and searching large amounts of data. Organizations can use Amazon ES to do the following:

- Analyze activity logs, e.g., logs for customer facing applications or websites
- Analyze CloudWatch logs with Elasticsearch
- Analyze product usage data coming from various services and systems
- Analyze social media sentiments, CRM data and find trends for your brand and products
- Analyze data stream updates from other AWS services, e.g., Amazon Kinesis Data Streams and Amazon DynamoDB
- Provide customers a rich search and navigation experience.

- Usage monitoring for mobile applications

Cost Model

With Amazon Elasticsearch Service, you pay only for what you use. There are no minimum fees or upfront commitments. You are charged for Amazon ES instance hours, Amazon EBS storage (if you choose this option), and standard data transfer fees.

You can get started with our free tier, which provides free usage of up to 750 hours per month of a single-AZ t2.micro.elasticsearch or t2.small.elasticsearch instance and 10 GB per month of optional Amazon EBS storage (Magnetic or General Purpose).

Amazon ES allows you to add data durability through automated and manual snapshots of your cluster. Amazon ES provides storage space for automated snapshots free of charge for each Amazon Elasticsearch Service domain. Automated snapshots are retained for a period of 14 days. Manual snapshots are charged according to Amazon S3 storage rates. Data transfer for using the snapshots is free of charge. For more information, see Amazon Elasticsearch Service Pricing.

Performance

Performance of Amazon ES depends on multiple factors including instance type, workload, index, number of shards used, read replicas, storage configurations –instance storage or EBS storage (general purpose SSD). Indexes are made up of shards of data which can be distributed on different instances in multiple Availability Zones.

Read replica of the shards are maintained by Amazon ES in a different Availability Zone if zone awareness is checked. Amazon ES can use either the fast SSD instance storage for storing indexes or multiple EBS volumes. A search engine makes heavy use of storage devices and making disks faster will result in faster query and search performance.

Durability and Availability

You can configure your Amazon ES domains for high availability by enabling the Zone Awareness option either at domain creation time or by modifying a live domain. When Zone Awareness is enabled, Amazon ES distributes the instances supporting the domain across two different Availability Zones. Then, if you enable replicas in Elasticsearch, the instances are automatically distributed in such a way as to deliver cross-zone replication. You can build data durability for your Amazon ES domain through automated and manual snapshots.

You can use snapshots to recover your domain with preloaded data or to create a new domain with preloaded data. Snapshots are stored in Amazon S3, which is a secure, durable, highly-scalable object storage. By default, Amazon ES automatically creates daily snapshots of each domain. In addition, you can use the Amazon ES snapshot APIs to create additional manual snapshots. The manual snapshots are stored in Amazon S3. Manual snapshots can be used for cross-region disaster recovery and to provide additional durability.

Scalability and Elasticity

You can add or remove instances, and easily modify Amazon EBS volumes to accommodate data growth. You can write a few lines of code that will monitor the state of your domain through Amazon CloudWatch metrics and call the Amazon Elasticsearch Service API to scale your domain up or down based on thresholds you set. The service will execute the scaling without any downtime. Amazon Elasticsearch Service supports 1 EBS volume (max size of 1.5 TB) per instance associated with a domain. With the default maximum of 20 data nodes allowed per Amazon ES domain, you can allocate about 30

TB of EBS storage to a single domain. You can request a service limit increase up to 100 instances per domain by creating a case with the AWS Support Center. With 100 instances, you can allocate about 150 TB of EBS storage to a single domain.

Interfaces

Amazon ES supports many of the commonly used Elasticsearch APIs, so code, applications, and popular tools that you're already using with your current Elasticsearch environments will work seamlessly. For a full list of supported Elasticsearch operations, see our documentation.

The AWS CLI, API, or the AWS Management Console can be used for creating and managing your domains as well.

Amazon ES supports integration with several AWS services, including streaming data from S3 buckets, Amazon Kinesis Data Streams, and DynamoDB Streams. Both integrations use a Lambda function as an event handler in the cloud that responds to new data in Amazon S3 and Amazon Kinesis Data Streams by processing it and streaming the data to your Amazon ES domain. Amazon ES also integrates with Amazon CloudWatch for monitoring Amazon ES domain metrics and CloudTrail for auditing configuration API calls to Amazon ES domains.

Amazon ES includes built-in integration with Kibana, an open-source analytics and visualization platform and supports integration with Logstash, an open-source data pipeline that helps you process logs and other event data. You can set up your Amazon ES domain as the backend store for all logs coming through your Logstash implementation to easily ingest structured and unstructured data from a variety of sources.

Anti-Patterns

- **Online transaction processing (OLTP)** - Amazon ES is a real-time distributed search and analytics engine. There is no support for transactions or processing on data manipulation. If your requirement is for a fast transactional system, then a traditional relational database system built on Amazon RDS, or a NoSQL database offering functionality such as DynamoDB, is a better choice.
- **Ad hoc data querying** – While Amazon ES takes care of the operational overhead of building a highly scalable Elasticsearch cluster, if running Ad hoc queries or one-off queries against your data set is your use-case, Amazon Athena is a better choice. Amazon Athena is an interactive query service that makes it easy to analyze data in Amazon S3 using standard SQL, without provisioning servers.

Amazon QuickSight

Amazon QuickSight is a very fast, easy-to-use, cloud-powered business analytics service that makes it easy for all employees within an organization to build visualizations, perform ad-hoc analysis, and quickly get business insights from their data, anytime, on any device. It can connect to a wide variety of data sources including flat files e.g. CSV and Excel, access on premise databases including SQL Server, MySQL and PostgreSQL., AWS resources like Amazon RDS databases, Amazon Redshift, Amazon Athena and Amazon S3. Amazon QuickSight enables organizations to scale their business analytics capabilities to hundreds of thousands of users, and delivers fast and responsive query performance by using a robust in-memory engine (SPICE).

Amazon QuickSight is built with "SPICE" – a Super-fast, Parallel, In-memory Calculation Engine. Built from the ground up for the cloud, SPICE uses a combination of columnar storage, in-memory technologies enabled through the latest hardware innovations and machine code generation to run interactive queries on large datasets and get rapid responses. SPICE supports rich calculations to help you derive valuable insights from your analysis without worrying about provisioning or managing infrastructure. Data in SPICE is persisted until it is explicitly deleted by the user. SPICE also automatically

replicates data for high availability and enables Amazon QuickSight to scale to hundreds of thousands of users who can all simultaneously perform fast interactive analysis across a wide variety of AWS data sources.

Topics

- Ideal Usage Patterns (p. 24)
- Cost Model (p. 24)
- Performance (p. 24)
- Durability and Availability (p. 25)
- Scalability and Elasticity (p. 25)
- Interfaces (p. 25)
- Anti-Patterns (p. 25)

Ideal Usage Patterns

Amazon QuickSight is an ideal Business Intelligence tool allowing end users to create visualizations that provide insight into their data to help them make better business decisions. Amazon QuickSight can be used to do the following:

- Quick interactive ad-hoc exploration and optimized visualization of data
- Create and share dashboards and KPI's to provide insight into your data
- Create Stories which are guided tours through specific views of an analysis and allow you to share insights and collaborate with others. They are used to convey key points, a thought process, or the evolution of an analysis for collaboration.
- Analyze and visualize data coming from logs and stored in S3
- Analyze and visualize data from on premise databases like SQL Server, Oracle, PostGreSQL, and MySQL
- Analyze and visualize data in various AWS resources, e.g., Amazon RDS databases, Amazon Redshift, Amazon Athena, and Amazon S3.
- Analyze and visualize data in software as a service (SaaS) applications like Salesforce.
- Analyze and visualize data in data sources that can be connected to using JDBC/ODBC connection.

Cost Model

Amazon QuickSight has two different editions for pricing; standard edition and enterprise edition. For an annual subscription it is $9/user/month for standard edition and $18/user/month for enterprise edition, both with 10 GB of SPICE capacity included. You can get additional SPICE capacity for $.25/GB/month for standard edition and $.38/GB/month for enterprise edition. We also have month to month option for both the editions. For standard edition it is $12/GB/month and enterprise edition is $24/GB/month. Additional information on pricing can be found at Amazon QuickSight Pricing.

Both editions offer a full set of features for creating and sharing data visualizations. Enterprise edition also offers encryption at rest and Microsoft Active Directory (AD) integration. In Enterprise edition, you select a Microsoft AD directory in AWS Directory Service. You use that active directory to identify and manage your Amazon QuickSight users and administrators.

Performance

Amazon QuickSight is built with 'SPICE', a Super-fast, Parallel, and In-memory Calculation Engine. Built from the ground up for the cloud, SPICE uses a combination of columnar storage, in-memory

technologies enabled through the latest hardware innovations, and machine code generation to run interactive queries on large datasets and get rapid responses.

Durability and Availability

SPICE automatically replicates data for high availability and enables Amazon QuickSight to scale to hundreds of thousands of users who can all simultaneously perform fast interactive analysis across a wide variety of AWS data sources.

Scalability and Elasticity

Amazon QuickSight is a fully managed service and it internally takes care of scaling to meet the demands of your end users. With Amazon QuickSight you don't need to worry about scale. You can seamlessly grow your data from a few hundred megabytes to many terabytes of data without managing any infrastructure.

Interfaces

Amazon QuickSight can connect to a wide variety of data sources including flat files (CSV, TSV, CLF, ELF), connect to on-premises databases like SQL Server, MySQL, and PostgreSQL, and AWS data sources including Amazon RDS, Amazon Aurora, Amazon Redshift, Amazon Athena and Amazon S3, and SaaS, applications like Salesforce. You can also export analyzes from a visual to a file with CSV format.

You can share an analysis, dashboard, or story using the share icon from the Amazon QuickSight service interface. You will be able to select the recipients (email address, username or group name), permission levels, and other options before sharing the content with others.

Anti-Patterns

- **Highly formatted canned Reports** – Amazon QuickSight is much more suited for ad hoc query, analysis and visualization of data. For highly formatted reports e.g. formatted financial statements consider using a different tool.
- **ETL** - While Amazon QuickSight can perform some transformations it is not a full-fledged ETL tool. AWS offers AWS Glue, which is a fully managed extract, transform, and load (ETL) service that makes it easy for customers to prepare and load their data for analytics.

Amazon EC2

Amazon EC2, with instances acting as AWS virtual machines, provides an ideal platform for operating your own self-managed big data analytics applications on AWS infrastructure. Almost any software you can install on Linux or Windows virtualized environments can be run on Amazon EC2 and you can use the pay-as-you-go pricing model. What you don't get are the application-level managed services that come with the other services mentioned in this whitepaper. There are many options for self-managed big data analytics; here are some examples:

- A NoSQL offering, such as MongoDB
- A data warehouse or columnar store like Vertica
- A Hadoop cluster
- An Apache Storm cluster
- An Apache Kafka environment

Topics

- Ideal Usage Patterns (p. 26)
- Cost Model (p. 26)
- Performance (p. 26)
- Durability and Availability (p. 26)
- Scalability and Elasticity (p. 26)
- Interfaces (p. 27)
- Anti-Patterns (p. 27)

Ideal Usage Patterns

- **Specialized Environment** – When running a custom application, a variation of a standard Hadoop set or an application not covered by one of our other offerings, Amazon EC2 provides the flexibility and scalability to meet your computing needs.
- **Compliance Requirements** – Certain compliance requirements may require you to run applications yourself on Amazon EC2 instead of using a managed service offering.

Cost Model

Amazon EC2 has a variety of instance types in a number of instance families (standard, high CPU, high memory, high I/O, etc.), and different pricing options (On-Demand, Reserved, and Spot). Depending on your application requirements, you may want to use additional services along with Amazon EC2, such as Amazon Elastic Block Store (Amazon EBS) for directly attached persistent storage or Amazon S3 as a durable object store; each comes with their own pricing model. If you do run your big data application on Amazon EC2, you are responsible for any license fees just as you would be in your own data center. The AWS Marketplace offers many different third-party, big data software packages preconfigured to launch with a simple click of a button.

Performance

Performance in Amazon EC2 is driven by the instance type that you choose for your big data platform. Each instance type has a different amount of CPU, RAM, storage, IOPs, and networking capability so that you can pick the right performance level for your application requirements.

Durability and Availability

Critical applications should be run in a cluster across multiple Availability Zones within an AWS Region so that any instance or data center failure does not affect application users. For non-uptime critical applications, you can back up your application to Amazon S3 and restore to any Availability Zone in the region if an instance or zone failure occurs. Other options exist, depending on which application you are running and the requirements, such as mirroring your application.

Scalability and Elasticity

Auto Scaling is a service that allows you to automatically scale your Amazon EC2 capacity up or down according to conditions that you define. With Auto Scaling, you can ensure that the number of EC2 instances you're using scales up seamlessly during demand spikes to maintain performance, and scales down automatically during demand lulls to minimize costs. Auto Scaling is particularly well suited for applications that experience hourly, daily, or weekly variability in usage. Auto Scaling is enabled by CloudWatch and available at no additional charge beyond CloudWatch fees.

Interfaces

Amazon EC2 can be managed programmatically via API, SDK, or the console. Metrics for compute utilization, memory utilization, storage utilization, network consumption, and read/write traffic to your instances are free of charge using the console or CloudWatch API operations.

The interfaces for your big data analytics software that you run on top of Amazon EC2 varies based on the characteristics of the software you choose.

Anti-Patterns

Amazon EC2 has the following anti-patterns:

- **Managed Service** – If your requirement is a managed service offering where you abstract the infrastructure layer and administration from the big data analytics, then this "do it yourself" model of managing your own analytics software on Amazon EC2 may not be the correct choice.
- **Lack of Expertise or Resources** – If your organization does not have, or does not want to expend, the resources or expertise to install and manage a high-availability installation for the system in question, you should consider using the AWS equivalent such as Amazon EMR, DynamoDB, Amazon Kinesis Data Streams, or Amazon Redshift.

Amazon Athena

Amazon Athena is an interactive query service that makes it easy to analyze data in Amazon S3 using standard SQL. Athena is serverless, so there is no infrastructure to setup or manage, and you can start analyzing data immediately. You don't need to load your data into Athena, as it works directly with data stored in S3. Just log into the Athena Console, define your table schema, and start querying. Amazon Athena uses Presto with full ANSI SQL support and works with a variety of standard data formats, including CSV, JSON, ORC, Apache Parquet, and Apache Avro.

Topics

- Ideal Usage Patterns (p. 27)
- Cost Model (p. 28)
- Performance (p. 28)
- Durability and Availability (p. 28)
- Scalability and Elasticity (p. 28)
- Security, Authorization and Encryption (p. 28)
- Interfaces (p. 29)
- Anti-Patterns (p. 29)

Ideal Usage Patterns

- **Interactive ad hoc querying for web logs** – Athena is a good tool for interactive one-time SQL queries against data on Amazon S3. For example, you could use Athena to run a query on web and application logs to troubleshoot a performance issue. You simply define a table for your data and start querying using standard SQL. Athena integrates with Amazon QuickSight for easy visualization.
- **To query staging data before loading into Redshift** – You can stage your raw data in S3 before processing and loading it into Redshift, and then use Athena to query that data.
- **Send AWS Service logs to S3 for Analysis with Athena –** CloudTrail, Cloudfront, ELB/ALB and VPC flow logs can be analyzed with Athena. AWS CloudTrail logs include details about any API calls made

to your AWS services, including from the console. CloudFront logs can be used to explore users' surfing patterns across web properties served by CloudFront. Querying ELB/ALB logs allows you to see the source of traffic, latency, and bytes transferred to and from Elastic Load Balancing instances and backend applications. VPC flow logs capture information about the IP traffic going to and from network interfaces in VPCs in the Amazon VPC service. The logs allow you to investigate network traffic patterns and identify threats and risks across your VPC estate.

- **Building Interactive Analytical Solutions with notebook-based solutions, e.g., RStudio, Jupyter, or Zeppelin -** Data scientists and Analysts are often concerned about managing the infrastructure behind big data platforms while running notebook-based solutions such as RStudio, Jupyter, and Zeppelin. Amazon Athena makes it easy to analyze data using standard SQL without the need to manage infrastructure. Integrating these notebook-based solutions with Amazon Athena gives data scientists a powerful platform for building interactive analytical solutions.

Cost Model

Amazon Athena has simple pay-as-you-go pricing, with no up-front costs or minimum fees, and you'll only pay for the resources you consume. It is priced per query, $5 per TB of data scanned, and charges based on the amount of data scanned by the query. You can save from 30% to 90% on your per-query costs and get better performance by compressing, partitioning, and converting your data into columnar formats. Converting data to the columnar format allows Athena to read only the columns it needs to process the query.

You are charged for the number of bytes scanned by Amazon Athena, rounded up to the nearest megabyte, with a 10 MB minimum per query. There are no charges for Data Definition Language (DDL) statements like CREATE/ALTER/DROP TABLE, statements for managing partitions, or failed queries. Cancelled queries are charged based on the amount of data scanned.

Performance

You can improve the performance of your query by compressing, partitioning, and converting your data into columnar formats. Amazon Athena supports open source columnar data formats such as Apache Parquet and Apache ORC. Converting your data into a compressed, columnar format lowers your cost and improves query performance by enabling Athena to scan less data from S3 when executing your query.

Durability and Availability

Amazon Athena is highly available and executes queries using compute resources across multiple facilities, automatically routing queries appropriately if a particular facility is unreachable. Athena uses Amazon S3 as its underlying data store, making your data highly available and durable. Amazon S3 provides durable infrastructure to store important data and is designed for durability of 99.999999999% of objects. Your data is redundantly stored across multiple facilities and multiple devices in each facility.

Scalability and Elasticity

Athena is serverless, so there is no infrastructure to setup or manage, and you can start analyzing data immediately. Since it is serverless it can scale automatically, as needed.

Security, Authorization and Encryption

Amazon Athena allows you to control access to your data by using AWS Identity and Access Management (IAM) policies, Access Control Lists (ACLs), and Amazon S3 bucket policies. With IAM policies, you can grant IAM users fine-grained control to your S3 buckets. By controlling access to data in S3, you can restrict users from querying it using Athena. You can query data that's been protected by:

- Server-side encryption with an Amazon S3-managed key
- Server-side encryption with an AWS KMS-managed key
- Client-side encryption with an AWS KMS-managed key

Amazon Athena also can directly integrate with AWS Key Management System (KMS) to encrypt your result sets, if desired.

Interfaces

Querying can be done by using the Athena Console. Athena also supports CLI, API via SDK and JDBC. Athena also integrates with Amazon QuickSight for creating visualizations based on the Athena queries.

Anti-Patterns

Amazon Athena has the following anti-patterns:

- **Enterprise Reporting and Business Intelligence Workloads** – Amazon Redshift is a better tool for Enterprise Reporting and Business Intelligence Workloads involving iceberg queries or cached data at the nodes. Data warehouses pull data from many sources, format and organize it, store it, and support complex, high speed queries that produce business reports. The query engine in Amazon Redshift has been optimized to perform especially well on data warehouse workloads.
- **ETL Workloads –** You should use Amazon EMR/Amazon Glue if you are looking for an ETL tool to process extremely large datasets and analyze them with the latest big data processing frameworks such as Spark, Hadoop, Presto, or Hbase.
- **RDBMS –** Athena is not a relational/transactional database. It is not meant to be a replacement for SQL engines like MySQL.

Solving Big Data Problems on AWS

In this whitepaper, we have examined some tools available on AWS for big data analytics. This paper provides a good reference point when starting to design your big data applications. However, there are additional aspects you should consider when selecting the right tools for your specific use case. In general, each analytical workload has certain characteristics and requirements that dictate which tool to use, such as:

- How quickly do you need analytic results: in real time, in seconds, or is an hour a more appropriate time frame?
- How much value will these analytics provide your organization and what budget constraints exist?
- How large is the data and what is its growth rate?
- How is the data structured?
- What integration capabilities do the producers and consumers have?
- How much latency is acceptable between the producers and consumers?
- What is the cost of downtime or how available and durable does the solution need to be?
- Is the analytic workload consistent or elastic?

Each one of these questions helps guide you to the right tool. In some cases, you can simply map your big data analytics workload into one of the services based on a set of requirements. However, in most real-world, big data analytic workloads, there are many different, and sometimes conflicting, characteristics and requirements on the same data set.

For example, some result sets may have real-time requirements as a user interacts with a system, while other analytics could be batched and run on a daily basis. These different requirements over the same data set should be decoupled and solved by using more than one tool. If you try to solve both of these examples using the same toolset, you end up either over-provisioning or therefore overpaying for unnecessary response time, or you have a solution that does not respond fast enough to your users in real time. Matching the best-suited tool to each analytical problem results in the most cost-effective use of your compute and storage resources.

Big data doesn't need to mean "big costs". So, when designing your applications, it's important to make sure that your design is cost efficient. If it's not, relative to the alternatives, then it's probably not the right design. Another common misconception is that using multiple tool sets to solve a big data problem is more expensive or harder to manage than using one big tool. If you take the same example of two different requirements on the same data set, the real-time request may be low on CPU but high on I/O, while the slower processing request may be very compute intensive.

Decoupling can end up being much less expensive and easier to manage because you can build each tool to exact specifications and not overprovision. With the AWS pay-as-you-go model, this equates to a much better value because you could run the batch analytics in just one hour and therefore only pay for the compute resources for that hour. Also, you may find this approach easier to manage rather than leveraging a single system that tries to meet all of the requirements. Solving for different requirements with one tool results in attempting to fit a square peg (real-time requests) into a round hole (a large data warehouse).

The AWS platform makes it easy to decouple your architecture by having different tools analyze the same data set. AWS services have built-in integration so that moving a subset of data from one tool to another can be done very easily and quickly using parallelization. Let's put this into practice by exploring a few real world, big data analytics problem scenarios and walking through an AWS architectural solution.

Topics

- Example 1: Queries against an Amazon S3 Data Lake (p. 31)

- Example 2: Capturing and Analyzing Sensor Data (p. 32)
- Example 3: Sentiment Analysis of Social Media (p. 34)

Example 1: Queries against an Amazon S3 Data Lake

Data lakes are an increasingly popular way to store and analyze both structured and unstructured data. If you use an Amazon S3 data lake, AWS Glue can make all your data immediately available for analytics without moving the data. AWS Glue crawlers can scan your data lake and keep the AWS AWS Glue Data Catalog in sync with the underlying data. You can then directly query your data lake with Amazon Athena and Amazon Redshift Spectrum. You can also use the AWS AWS Glue Data Catalog as your external Apache Hive Metastore for big data applications running on Amazon EMR.

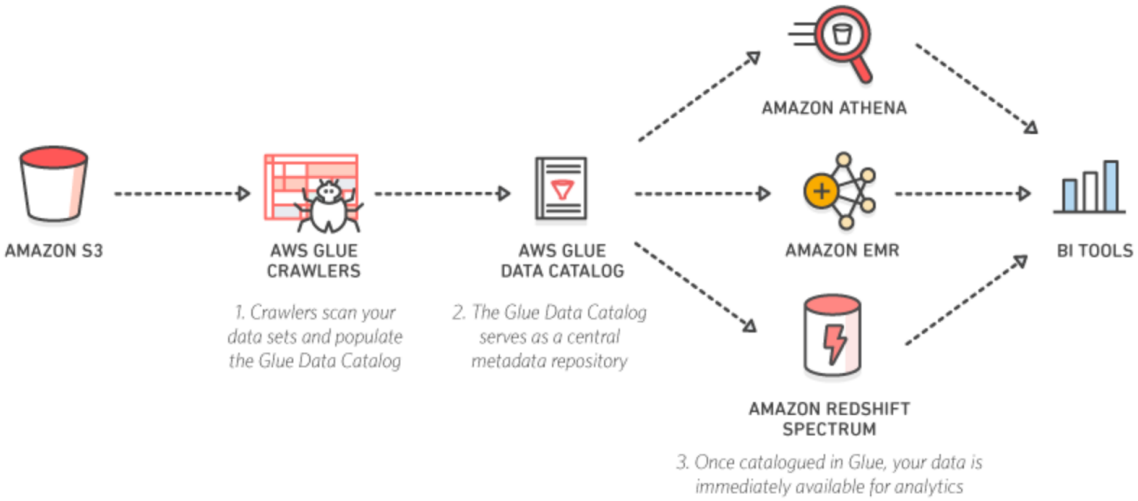

Figure 1: Queries against an Amazon S3 Data Lake

1. An AWS Glue crawler connects to a data store, progresses through a prioritized list of classifiers to extract the schema of your data and other statistics, and then populates the AWS AWS Glue Data Catalog with this metadata. Crawlers can run periodically to detect the availability of new data as well as changes to existing data, including table definition changes. Crawlers automatically add new tables, new partitions to existing table, and new versions of table definitions. You can customize AWS Glue crawlers to classify your own file types.
2. The AWS AWS Glue Data Catalog is a central repository to store structural and operational metadata for all your data assets. For a given data set, you can store its table definition, physical location, add business relevant attributes, as well as track how this data has changed over time. The AWS AWS Glue Data Catalog is Apache Hive Metastore compatible and is a drop-in replacement for the Apache Hive Metastore for Big Data applications running on Amazon EMR. For more information on setting up your EMR cluster to use AWS AWS Glue Data Catalog as an Apache Hive Metastore, click here.
3. The AWS AWS Glue Data Catalog also provides out-of-box integration with Amazon Athena, Amazon EMR, and Amazon Redshift Spectrum. Once you add your table definitions to the AWS AWS Glue Data Catalog, they are available for ETL and also readily available for querying in Amazon Athena, Amazon EMR, and Amazon Redshift Spectrum so that you can have a common view of your data between these services.
4. Using a BI tool like Amazon QuickSight enables you to easily build visualizations, perform ad-hoc analysis, and quickly get business insights from your data. Amazon QuickSight supports data sources

like: Amazon Athena, Amazon Redshift Spectrum, Amazon S3 and many others, see here: Supported Data Sources.

Example 2: Capturing and Analyzing Sensor Data

An international air conditioner manufacturer has many large air conditioners that it sells to various commercial and industrial companies. Not only do they sell the air conditioner units but, to better position themselves against their competitors, they also offer add-on services where you can see real-time dashboards in a mobile app or a web browser. Each unit sends its sensor information for processing and analysis. This data is used by the manufacturer and its customers. With this capability, the manufacturer can visualize the dataset and spot trends.

Currently, they have a few thousand pre-purchased air conditioning (A/C) units with this capability. They expect to deliver these to customers in the next couple of months and are hoping that, in time, thousands of units throughout the world will be using this platform. If successful, they would like to expand this offering to their consumer line as well, with a much larger volume and a greater market share. The solution needs to be able to handle massive amounts of data and scale as they grow their business without interruption. How should you design such a system?

First, break it up into two work streams, both originating from the same data:

- A/C unit's current information with near real-time requirements and a large number of customers consuming this information.
- All historical information on the A/C units to run trending and analytics for internal use.

The data-flow architecture in the following illustration shows how to solve this big data problem.

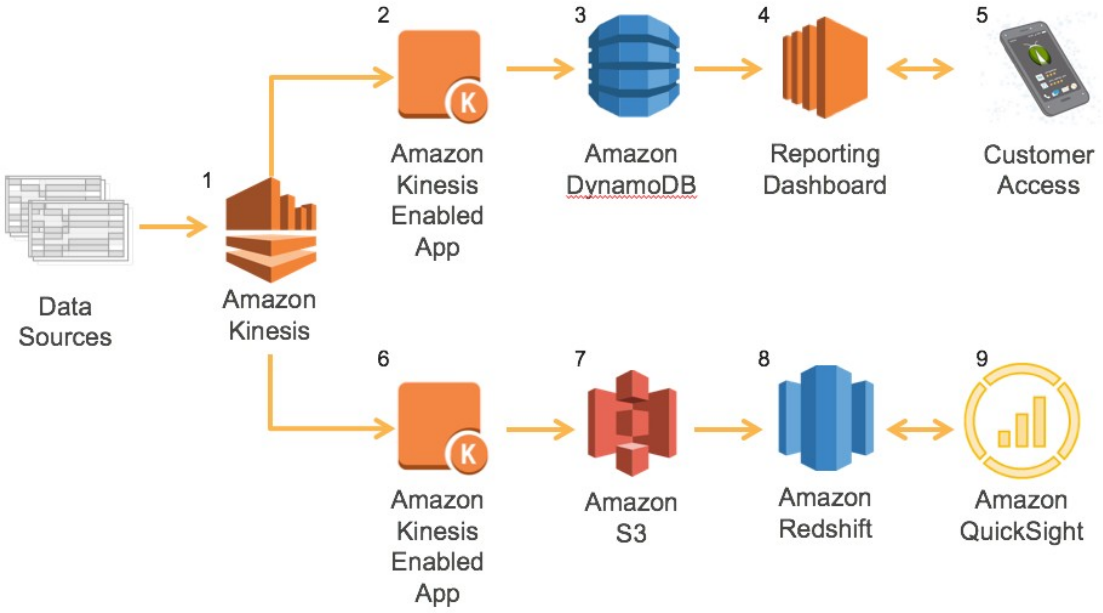

Figure 2: Capturing and Analyzing Sensor Data

1. The process begins with each A/C unit providing a constant data stream to Amazon Kinesis Data Streams. This provides an elastic and durable interface the units can talk to that can be scaled seamlessly as more and more A/C units are sold and brought online.

2. Using the Amazon Kinesis Data Streams-provided tools such as the Kinesis Client Library or SDK, a simple application is built on Amazon EC2 to read data as it comes into Amazon Kinesis Data Streams, analyze it, and determine if the data warrants an update to the real-time dashboard. It looks for changes in system operation, temperature fluctuations, and any errors that the units encounter.
3. This data flow needs to occur in near real time so that customers and maintenance teams can be alerted as quickly as possible if there is an issue with the unit. The data in the dashboard does have some aggregated trend information, but it is mainly the current state as well as any system errors. So, the data needed to populate the dashboard is relatively small. Additionally, there will be lots of potential access to this data from the following sources:

- Customers checking on their system via a mobile device or browser
- Maintenance teams checking the status of its fleet
- Data and intelligence algorithms and analytics in the reporting platform spot trends that can be then sent out as alerts, such as if the A/C fan has been running unusually long with the building temperature not going down.

DynamoDB was chosen to store this near real-time data set because it is both highly available and scalable; throughput to this data can be easily scaled up or down to meet the needs of its consumers as the platform is adopted and usage grows.

4. The reporting dashboard is a custom web application that is built on top of this data set and run on Amazon EC2. It provides content based on the system status and trends as well as alerting customers and maintenance crews of any issues that may come up with the unit.
5. The customer accesses the data from a mobile device or a web browser to get the current status of the system and visualize historical trends.

The data flow (steps 2-5) that was just described is built for near real-time reporting of information to human consumers. It is built and designed for low latency and can scale very quickly to meet demand. The data flow (steps 6-9) that is depicted in the lower part of the diagram does not have such stringent speed and latency requirements. This allows the architect to design a different solution stack that can hold larger amounts of data at a much smaller cost per byte of information and choose less expensive compute and storage resources.

6. To read from the Amazon Kinesis stream, there is a separate Amazon Kinesis-enabled application that probably runs on a smaller EC2 instance that scales at a slower rate. While this application is going to analyze the same data set as the upper data flow, the ultimate purpose of this data is to store it for long-term record and to host the data set in a data warehouse. This data set ends up being all data sent from the systems and allows a much broader set of analytics to be performed without the near real-time requirements.
7. The data is transformed by the Amazon Kinesis-enabled application into a format that is suitable for long-term storage, for loading into its data warehouse, and storing on Amazon S3. The data on Amazon S3 not only serves as a parallel ingestion point to Amazon Redshift, but is durable storage that will hold all data that ever runs through this system; it can be the single source of truth. It can be used to load other analytical tools if additional requirements arise. Amazon S3 also comes with native integration with Amazon Glacier, if any data needs to be cycled into long-term, low-cost storage.
8. Amazon Redshift is again used as the data warehouse for the larger data set. It can scale easily when the data set grows larger, by adding another node to the cluster.
9. For visualizing the analytics, one of the many partner visualization platforms can be used via the OBDC/JDBC connection to Amazon Redshift. This is where the reports, graphs, and ad hoc analytics can be performed on the data set to find certain variables and trends that can lead to A/C units underperforming or breaking.

This architecture can start off small and grow as needed. Additionally, by decoupling the two different work streams from each other, they can grow at their own rate without upfront commitment, allowing the manufacturer to assess the viability of this new offering without a large initial investment. You could easily imagine further additions, such as adding Amazon ML to predict how long an A/C unit will last and pre-emptively sending out maintenance teams based on its prediction algorithms to give their customers the best possible service and experience. This level of service would be a differentiator to the competition and lead to increased future sales.

Example 3: Sentiment Analysis of Social Media

A large toy maker has been growing very quickly and expanding their product line. After each new toy release, the company wants to understand how consumers are enjoying and using their products. Additionally, the company wants to ensure that their consumers are having a good experience with their products. As the toy ecosystem grows, the company wants to ensure that their products are still relevant to their customers and that they can plan future roadmaps items based on customer feedback. The company wants to capture the following insights from social media:

- Understand how consumers are using their products
- Ensure customer satisfaction
- Plan future roadmaps

Capturing the data from various social networks is relatively easy but the challenge is building the intelligence programmatically. After the data is ingested, the company wants to be able to analyze and classify the data in a cost-effective and programmatic way. To do this, you can use the architecture in the following illustration.

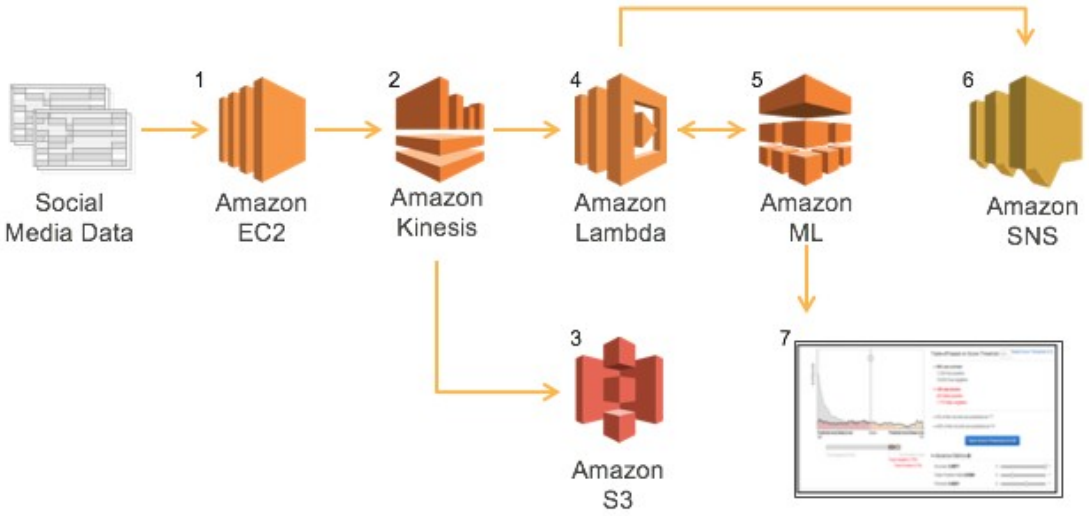

Figure 3: Sentiment Analysis of Social Media

The first step is to decide which social media sites to listen to. Then create an application on Amazon EC2 that polls those sites using their corresponding APIs.

Next, create an Amazon Kinesis stream, because we might have multiple data sources: Twitter, Tumblr, and so on. This way, a new stream can be created each time a new data source is added and you can take advantage of the existing application code and architecture. In this example, a new Amazon Kinesis stream is created to copy the raw data to Amazon S3 as well.

For archival, long term analysis, and historical reference, raw data is stored into Amazon S3. Additional Amazon ML batch models can be run on the data in Amazon S3 to perform predictive analysis and track consumer buying trends.

As noted in the architecture diagram, Lambda is used for processing and normalizing the data and requesting predictions from Amazon ML. After the Amazon ML prediction is returned, the Lambda function can take actions based on the prediction – for example, to route a social media post to the customer service team for further review.

Amazon ML is used to make predictions on the input data. For example, an ML model can be built to analyze a social media comment to determine whether the customer expressed negative sentiment about a product. To get accurate predictions with Amazon ML, start with training data and ensure that your ML models are working properly. If you are creating ML models for the first time, see Tutorial: Using Amazon ML to Predict Responses to a Marketing Offer. As mentioned earlier, if multiple social network data sources are used, then a different ML model for each one is suggested to ensure prediction accuracy.

Finally, actionable data is sent to Amazon SNS using Lambda, and delivered to the proper resources by text message or email for further investigation.

As part of the sentiment analysis, creating an Amazon ML model that is updated regularly is imperative for accurate results. Additional metrics about a specific model can be graphically displayed via the console, such as: accuracy, false positive rate, precision, and recall. For more information, see Step 4: Review the ML Model Predictive Performance and Set a Cut-Off.

By using a combination of Amazon Kinesis Data Streams, Lambda, Amazon ML, and Amazon SES, we have created a scalable and easily customizable social listening platform. Note that this scenario does not describe creating an Amazon ML model. You would create the model initially and then need to update it periodically, or as workloads change, to keep it accurate.

Conclusion

As more and more data is generated and collected, data analysis requires scalable, flexible, and high performing tools to provide insights in a timely fashion. However, organizations are facing a growing big data ecosystem where new tools emerge and “die” very quickly. Therefore, it can be very difficult to keep pace and choose the right tools.

This whitepaper offers a first step to help you solve this challenge. With a broad set of managed services to collect, process, and analyze big data, the AWS platform makes it easier to build, deploy, and scale big data applications. This allows you to focus on business problems instead of updating and managing these tools.

AWS provides many solutions to address your big data analytic requirements. Most big data architecture solutions use multiple AWS tools to build a complete solution. This approach helps meet stringent business requirements in the most cost-optimized, performant, and resilient way possible. The result is a flexible, big data architecture that is able to scale along with your business.

Contributors

The following individuals and organizations contributed to this document:

- Erik Swensson, Manager, Solutions Architecture, Amazon Web Services
- Erick Dame, Solutions Architect, Amazon Web Services
- Shree Kenghe, Solutions Architect, Amazon Web Services

Further Reading

The following resources can help you get started in running big data analytics on AWS:

- Big Data on AWS

View the comprehensive portfolio of big data services as well as links to other resources such AWS big data partners, tutorials, articles, and AWS Marketplace offerings on big data solutions. Contact us if you need any help.

- Read the AWS Big Data Blog

The blog features real life examples and ideas updated regularly to help you collect, store, clean, process, and visualize big data.

- Try one of the Big Data Test Drives

Explore the rich ecosystem of products designed to address big data challenges using AWS. Test Drives are developed by AWS Partner Network (APN) Consulting and Technology partners and are provided free of charge for education, demonstration, and evaluation purposes.

- Take an AWS training course on big data

The Big Data on AWS course introduces you to cloud-based big data solutions and Amazon EMR. We show you how to use Amazon EMR to process data using the broad ecosystem of Hadoop tools like Pig and Hive. We also teach you how to create big data environments, work with DynamoDB and Amazon Redshift, understand the benefits of Amazon Kinesis Streams, and leverage best practices to design big data environments for security and cost-effectiveness.

- View the Big Data Customer Case Studies

Learn from the experience of other customers who have built powerful and successful big data platforms on the AWS cloud.

Document Revisions

To be notified about updates to this whitepaper, subscribe to the RSS feed.

update-history-change	update-history-description	update-history-date
Whitepaper updated (p. 39)	Revised to add information on Amazon Athena, AWS QuickSight, AWS Glue, and general updates throughout.	December 1, 2018
Whitepaper updated (p. 39)	Revised to add information on Amazon Machine Learning, AWS Lambda, Amazon Elasticsearch Service; general update.	January 1, 2016
Initial publication (p. 39)	Whitepaper first published.	December 1, 2014

Notices

Customers are responsible for making their own independent assessment of the information in this document. This document: (a) is for informational purposes only, (b) represents current AWS product offerings and practices, which are subject to change without notice, and (c) does not create any commitments or assurances from AWS and its affiliates, suppliers or licensors. AWS products or services are provided “as is” without warranties, representations, or conditions of any kind, whether express or implied. The responsibilities and liabilities of AWS to its customers are controlled by AWS agreements, and this document is not part of, nor does it modify, any agreement between AWS and its customers.