
AWS
re:Invent

ANT204

An introduction to data lakes and analytics on AWS

Francis Jayakumar
Senior Analytics Specialist
AWS

Anna Coniglio
Senior Data Warehouse Solutions Architect
AWS

AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Epic Games continually improves Fortnite for 250+ million players globally

Challenge

They needed a way to process and analyze over 100PB of data (125M events / min.) ingested from game clients and game servers to understand and adapt to player engagement

Solution

Epic Games turned to AWS for an Amazon S3 data lake in combination with Amazon Elastic MapReduce (EMR), Amazon Elastic Compute Cloud (EC2), and Amazon Kinesis

Benefits

The data provides a constant feedback loop for designers and an up-to-the-minute analysis of gamer satisfaction to drive gamer engagement

EPIC GAMES

Customers want more value from their data



Growing
exponentially



From
new sources



Increasingly
diverse

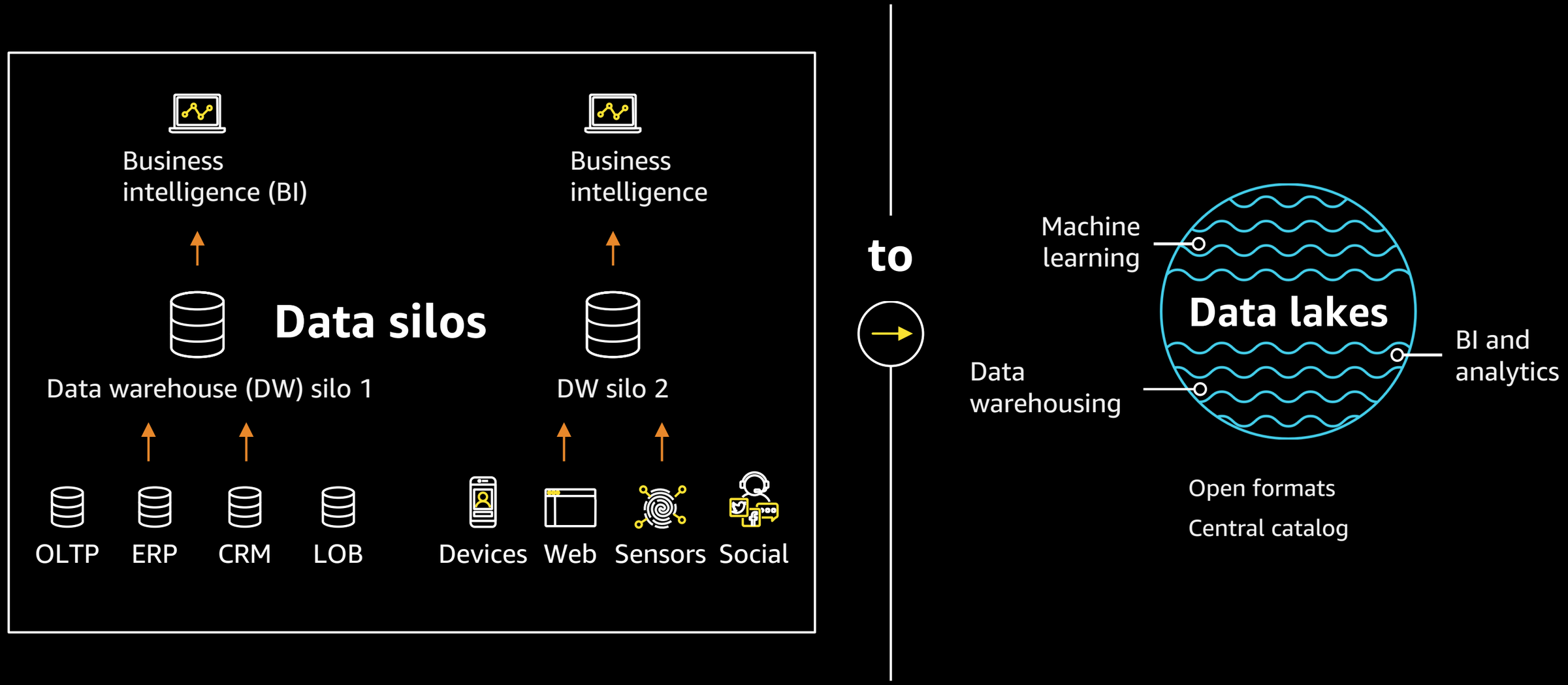


Used by
many people

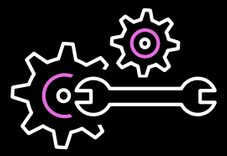


Analyzed by
many applications

Traditional data warehousing approaches don't scale



Why choose AWS for data lakes and analytics?



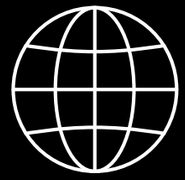
Easiest-to-build data lakes and analytics

1



Most secure infrastructure for analytics

2



Most comprehensive and open

3



Most scalable and cost effective

4

More data lakes and analytics than anywhere else

TENS OF THOUSANDS OF DATA LAKES RUN ON AWS ACROSS ALL INDUSTRIES



The AWS analytics portfolio

Data, visualization, engagement, and machine learning


AWS Data Exchange


Amazon QuickSight


Amazon Pinpoint


Amazon SageMaker


Amazon Comprehend


Amazon Lex


Amazon Polly


Amazon Rekognition


Amazon Translate

+ many more

Analytics


Amazon Redshift


Amazon EMR (Spark and Hadoop)


AWS Glue (Spark and Python)


Amazon Athena


Amazon Elasticsearch Service


Amazon Kinesis Data Analytics

Data lake infrastructure and management


Amazon S3 / Amazon S3 Glacier


AWS Lake Formation


AWS Glue

Data movement

Database migration service | Snowball | Snowmobile | Kinesis Data Streams | Kinesis Data Firehose | Managed streaming for Apache Kafka

Data movement services

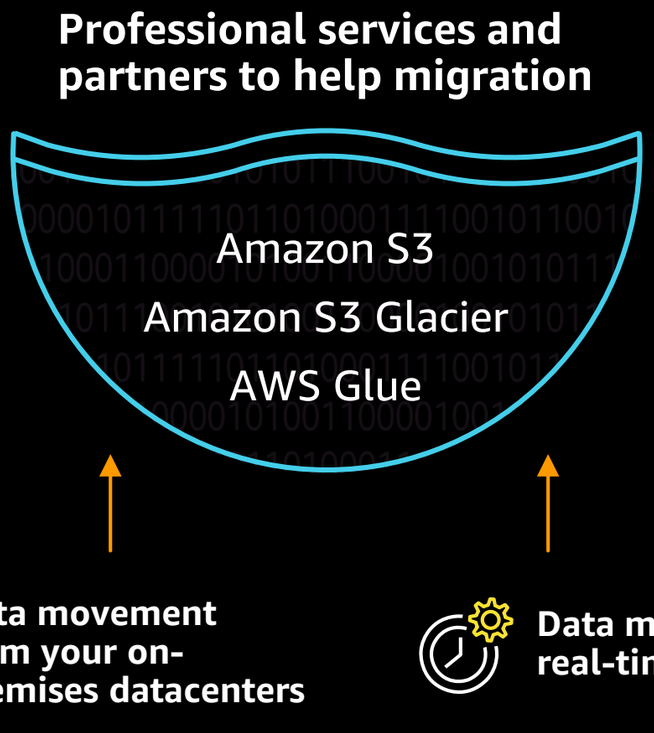
Data movement

Migration and streaming services

AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Most ways to move data to the data lake



Data movement from on-premises datacenters

Dedicated network connection

Secure appliances

Ruggedized shipping containers

Database migration

Gateway that lets applications write to the cloud

Data movement from real-time sources

Connect devices to AWS

Real-time data streams

Real-time video streams

Data lake infrastructure and management services

Data lake infrastructure and management



Amazon S3 / Amazon S3 Glacier



AWS Lake Formation



AWS Glue

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS Glue

SIMPLE, FLEXIBLE, AND COST-EFFECTIVE ETL AND DATA CATALOG

Data lake infrastructure and management

Less hassle



Integrated across AWS – Supports Amazon Aurora, Amazon RDS, Amazon Redshift, S3, and common database engines in your VPC running on EC2

Serverless



Serverless – No infrastructure to provision or manage

More power



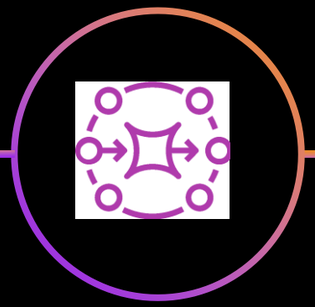
Automatically generates the code to execute your data transformations and loading processes

AWS Glue DataBrew

VISUAL DATA PREPARATION

Data lake infrastructure and management

Built-in transformations



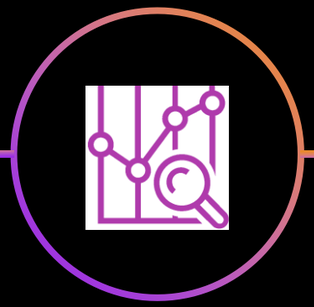
Join, tokenize, split, merge, extract, remove, group, pivot, normalize, one hot or label encode, and more

Serverless



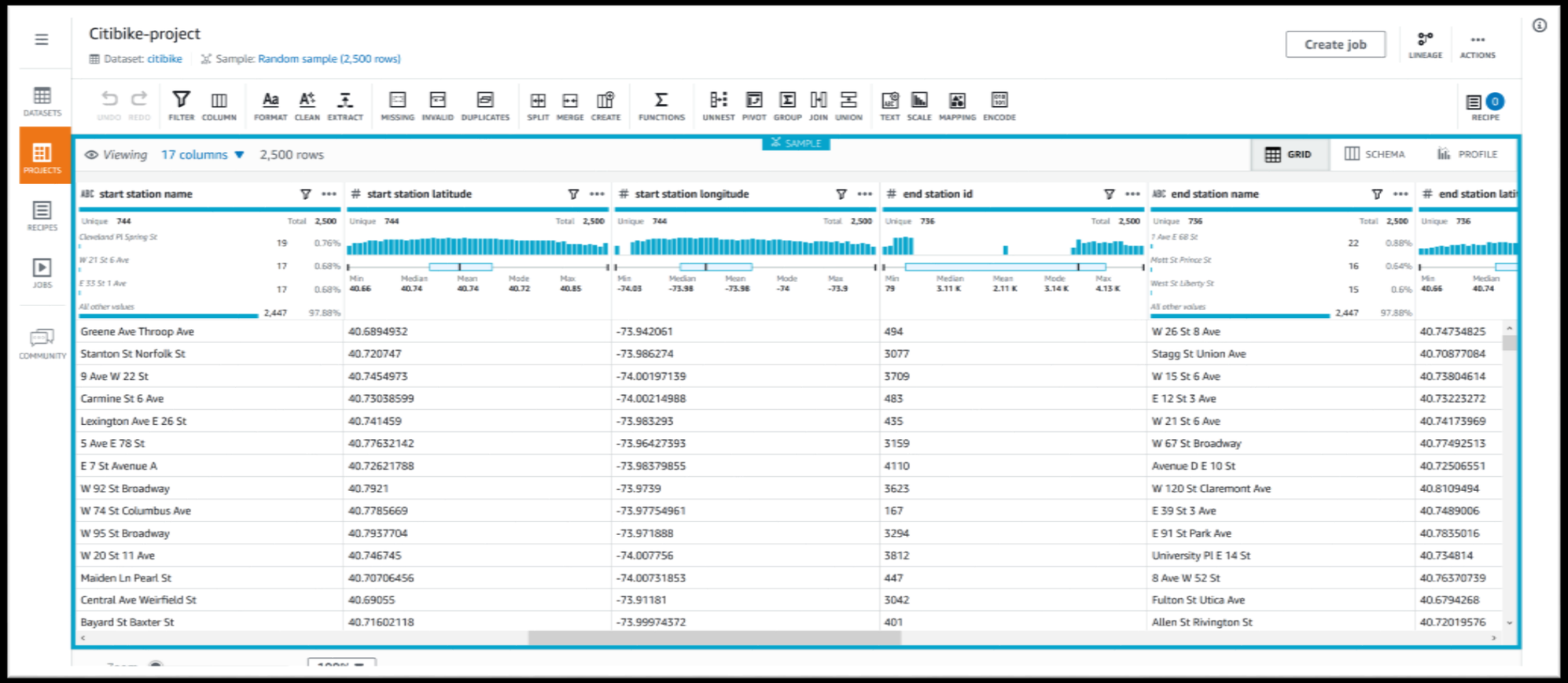
No infrastructure to provision or manage; pay only for what you use

For users of all technical levels



Visually interact with your data in a project as you build recipes

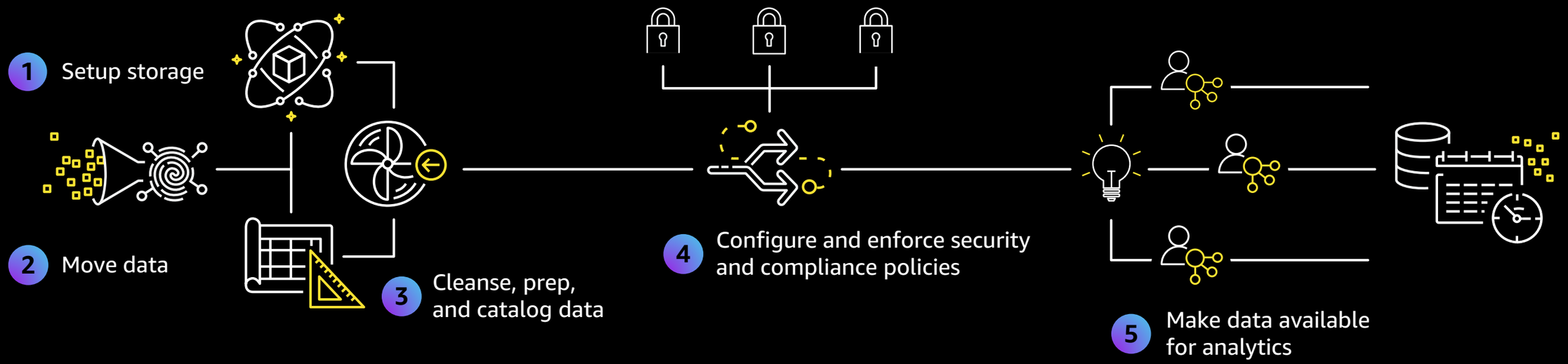
AWS Glue DataBrew: Visual data preparation



Challenges to making a secure data lake

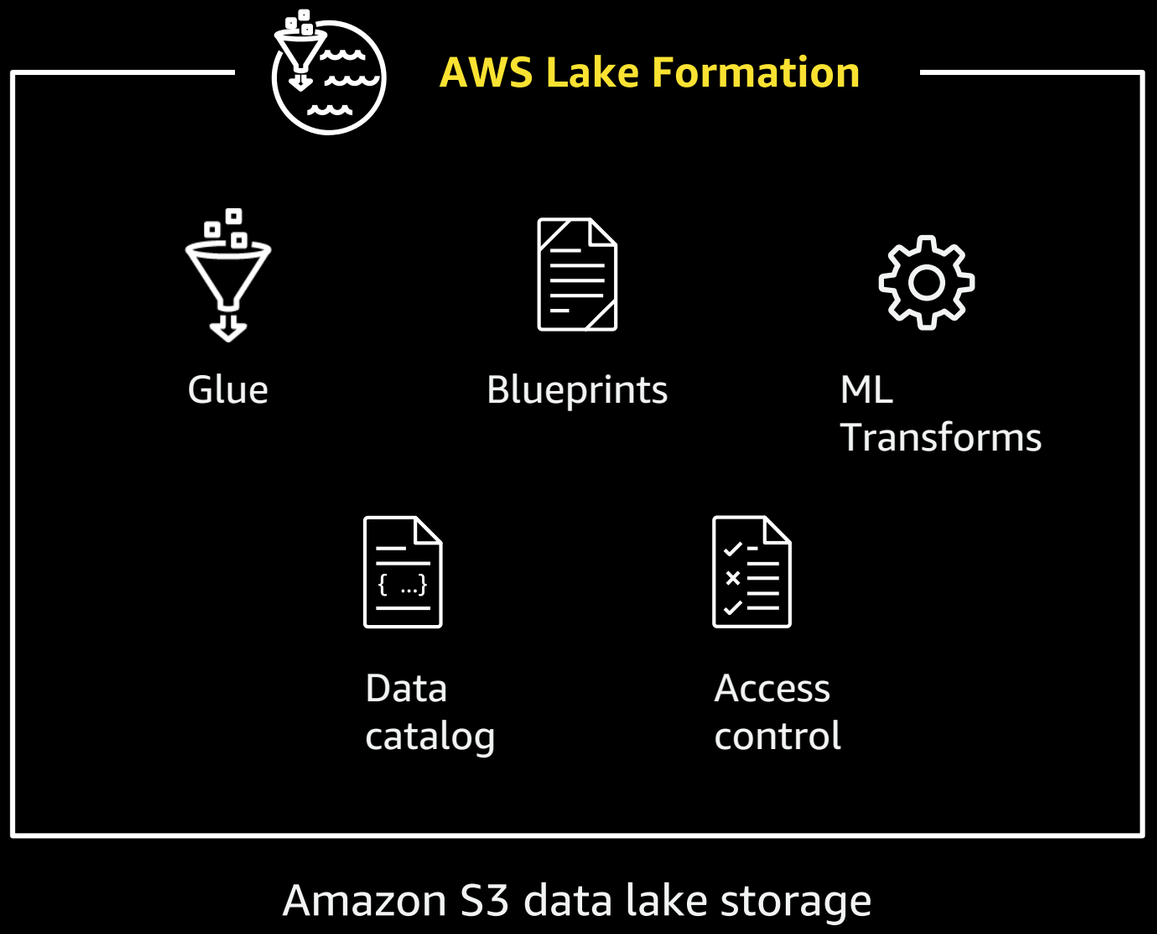
Data lake infrastructure and management

Typical steps of building a data lake



AWS Lake Formation

BUILD A SECURE DATA LAKE IN DAYS



- Comprehensive set of integrated tools enable user access consistently
- Centralized management of fine-grained permissions empower security officers
- Simplified ingest and cleaning enables data engineers to build faster

Analytics services

Analytics


Data warehousing


Big data processing


Serverless data processing


Interactive query


Operational analytics


Real-time analytics

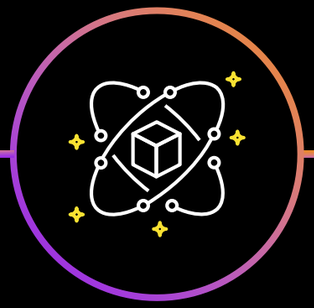
© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon EMR

Analytics

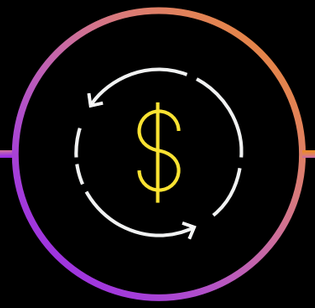
EASILY RUN SPARK, HADOOP, HIVE, PRESTO, HBASE, AND MORE BIG DATA APPS ON AWS

Latest versions



Updated with latest open-source frameworks within 30 days

Low cost



50–80% reduction in costs with EC2 Spot and reserved instances

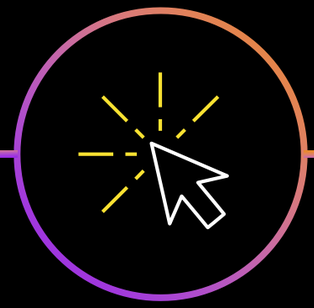
Per-second billing for flexibility

Use S3 storage



Process data in S3 securely with high performance using the EMR File System (EMRFS) connector

Easy



Fully managed no cluster setup, node provisioning, cluster tuning

Data warehousing: Amazon Redshift

FIRST AND MOST POPULAR CLOUD DATA WAREHOUSE

Data lake and AWS integration



Analyze exabytes of data across data warehouse, data lakes, and operational database

Query data across various analytics services

Best performance, most scalable



3x faster with RA3*

10x faster with AQUA*

Adds compute capacity on demand to meet unlimited concurrent access

*Vs. other cloud DWs

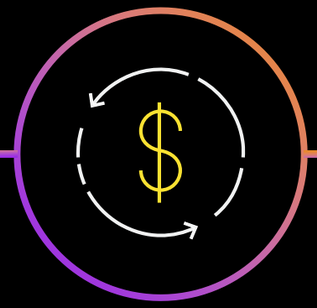
Most secure and compliant



AWS-grade security (e.g., VPC, encryption with KMS, CloudTrail)

All major certifications such as SOC, PCI, DSS, ISO, FedRAMP, HIPAA

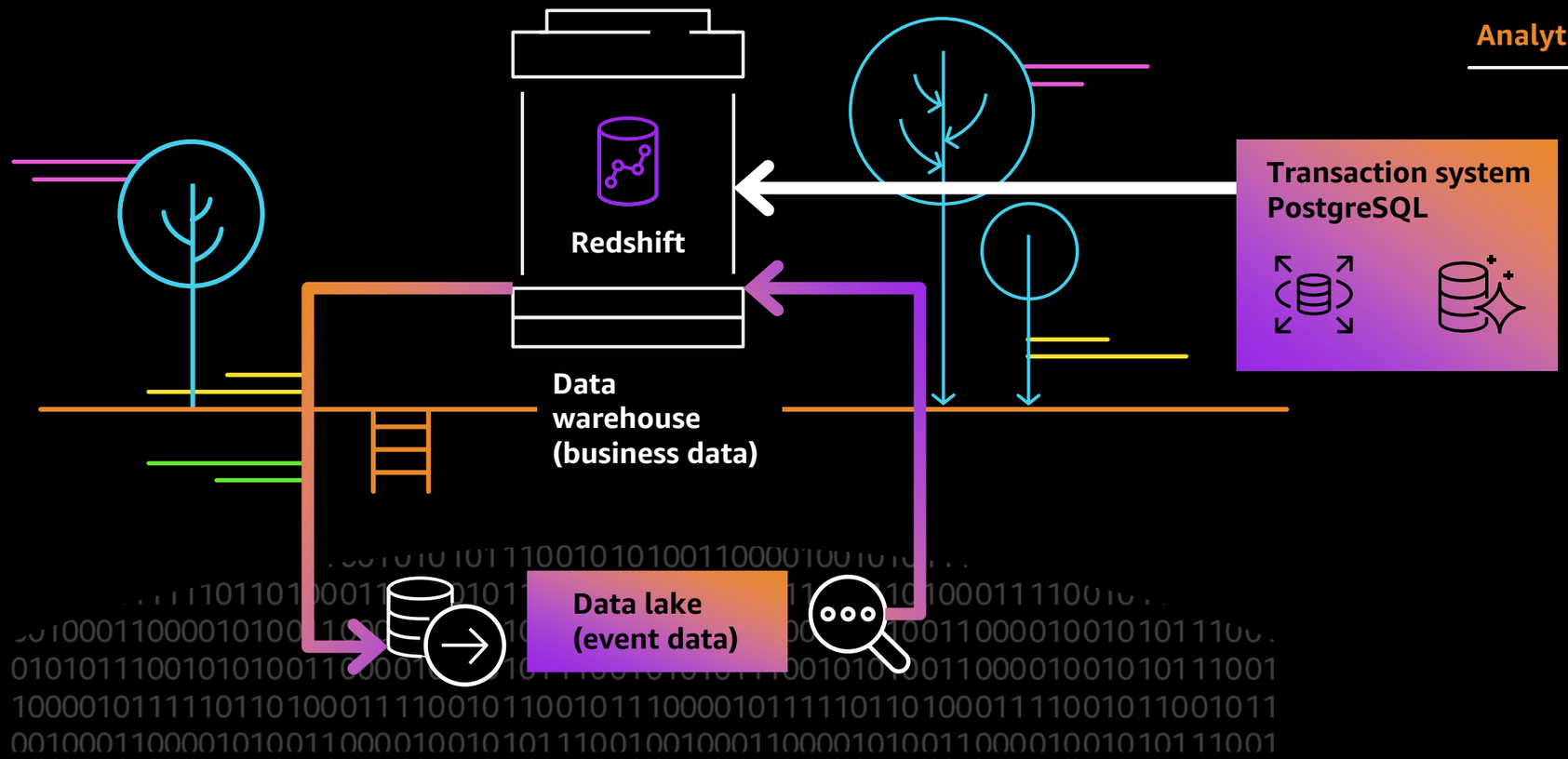
Lowest cost



Cost-optimized workloads by paying compute and storage separately

1/10th the cost of traditional DW at $1000 / TB / year

Up to 75% less than other cloud data warehouses and predictable costs



Customers moving to data lake architectures

Amazon Redshift enables you to have a lake house approach

Real time: Amazon Kinesis

Analytics

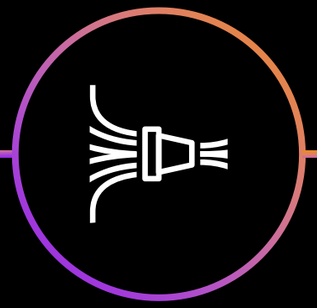
EASILY COLLECT, PROCESS, AND ANALYZE DATA AND VIDEO STREAMS IN REAL TIME

Kinesis Data Streams



Collect streaming data, at scale, for real-time analytics

Kinesis Data Firehose



Prepare and load data streams into data stores and analytics services

Kinesis Data Analytics



Get actionable insights from streaming data in real time

Kinesis Video Streams

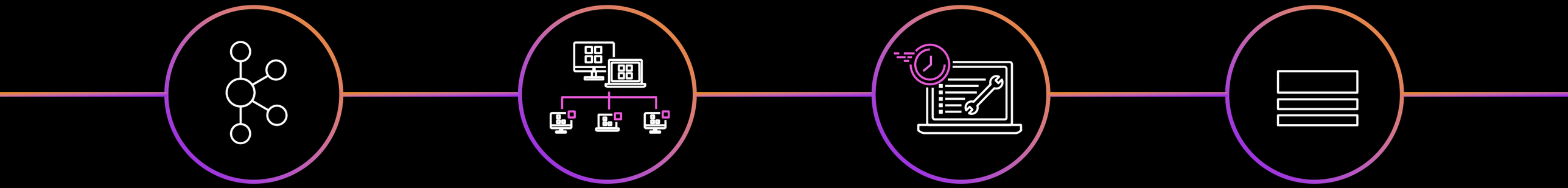


Capture, process, and store media streams for playback, analytics, and machine learning

Amazon-managed streaming for Apache Kafka

Analytics

FULLY-MANAGED, HIGHLY AVAILABLE, AND SECURE APACHE KAFKA SERVICE



Runs and manages Apache Kafka for you

Fully managed

Highly available

Multiple levels of security

Operational analytics

FULLY MANAGED, SCALABLE, SECURE, ELASTICSEARCH SERVICE

Open source Elasticsearch application programming interfaces (APIs), Kibana, and Logstash



Open-source Elasticsearch APIs

Managed Kibana

Integration with Logstash

Fully managed



Deploy Elasticsearch clusters in minutes – Simplified hardware provisioning, software installation / patching, failure recovery, backups, and monitoring

Scalable, secure, and compliant



Scale clusters up / down via a single API call or a few clicks

Secured network isolation with VPC, encrypt data at rest and in transit

Compliant: HIPAA, PCI DSS, and ISO

Pay only for what you use



Cost-optimized workloads

No upfront fee or usage requirement

Critical features built in – Encryption, VPC support, 24 / 7 monitoring

Amazon Athena

SERVERLESS, INTERACTIVE QUERY SERVICE

Analytics

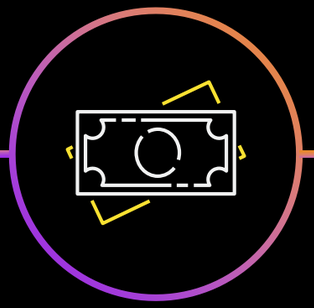
Query instantly



Zero setup cost

Point to S3 and start querying

Pay per query



Pay only for queries run

Save 30–90% on per-query costs through compression

Use S3 storage

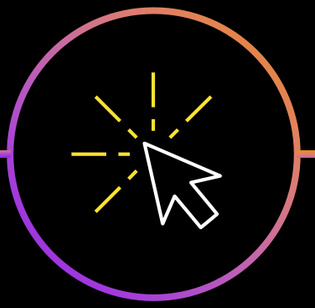


ANSI SQL

JDBC / ODBC drivers

Multiple formats, compression types, and complex joins and data types

Easy



Serverless – Zero infrastructure, zero administration

Integrated with QuickSight

Serverless analytics

DELIVER ON-DEMAND ANALYTICS ON THE DATA LAKE



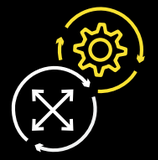
Serverless

Zero infrastructure

Zero administration



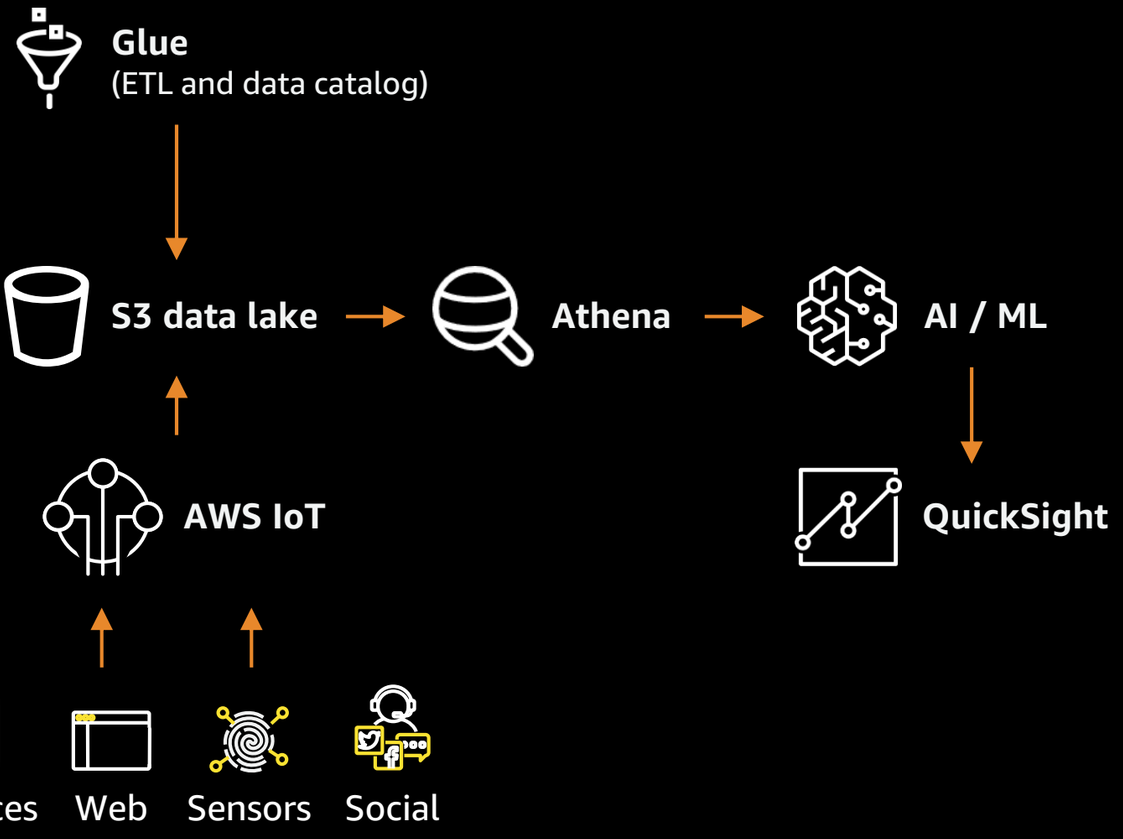
Never pay for idle resources



Automatically scales resources with usage



Availability and fault tolerance built in



Data, visualization, engagement, and machine learning services

Data, visualization, engagement, and machine learning



Data



Dashboards



Digital user engagement



Predictive analytics

AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS Data Exchange

EASILY FIND AND SUBSCRIBE TO THIRD-PARTY DATA IN THE CLOUD

Data, visualization, engagement, and ML

Quickly find diverse data in one place



>1,000 data products

>80 data providers including Dow Jones, Change Healthcare, Foursquare, Dun and Bradstreet, Thomson Reuters, Pitney Bowes, Lexis Nexis, and Deloitte

Easily analyze data

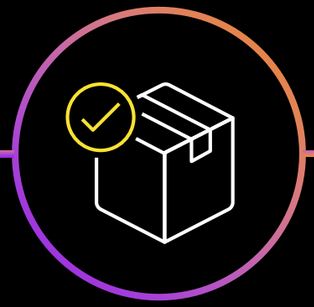


Download or copy data to S3

Combine, analyze, and model with existing data

Analyze data with EMR, Amazon Redshift, Athena, and AWS Glue

Efficiently access third-party data



Simplifies access to data – No need to receive physical media, manage FTP credentials, or integrate with different APIs

Minimize legal reviews and negotiations

Amazon QuickSight

FIRST BI SERVICE BUILT FOR THE CLOUD WITH PAY-PER-SESSION PRICING AND ML INSIGHTS

Data, visualization, engagement, and ML

Elastic scaling



Auto-scale 10 to 10K+ users in minutes

Pay as you go

Serverless



Create dashboards in minutes

Deploy globally without provisioning a single server

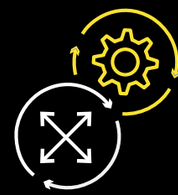
Deeply integrated with AWS services



Secure, private access to AWS data

Integrated S3 data lake permissions through AWS IAM

API support



Programmatically onboard users and manage content

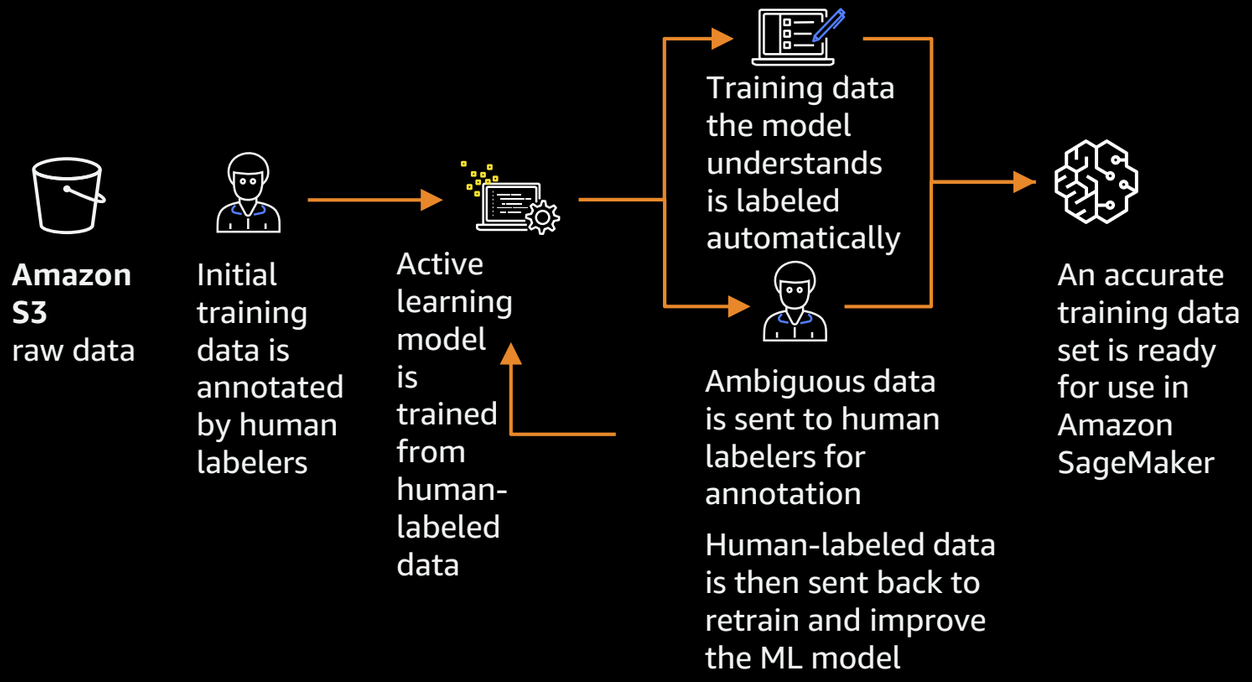
Easily embed in your apps

ML and AI services

PREDICTIVE INSIGHTS

Data, visualization, engagement, and ML

- Artificial intelligence (AI) services that enable developers to plug in pre-built AI functionality into their apps
- ML platform services that make it easy for any developer to get started and get deep with ML
- ML frameworks and interfaces for ML practitioners



Demonstration



AWS re:Invent

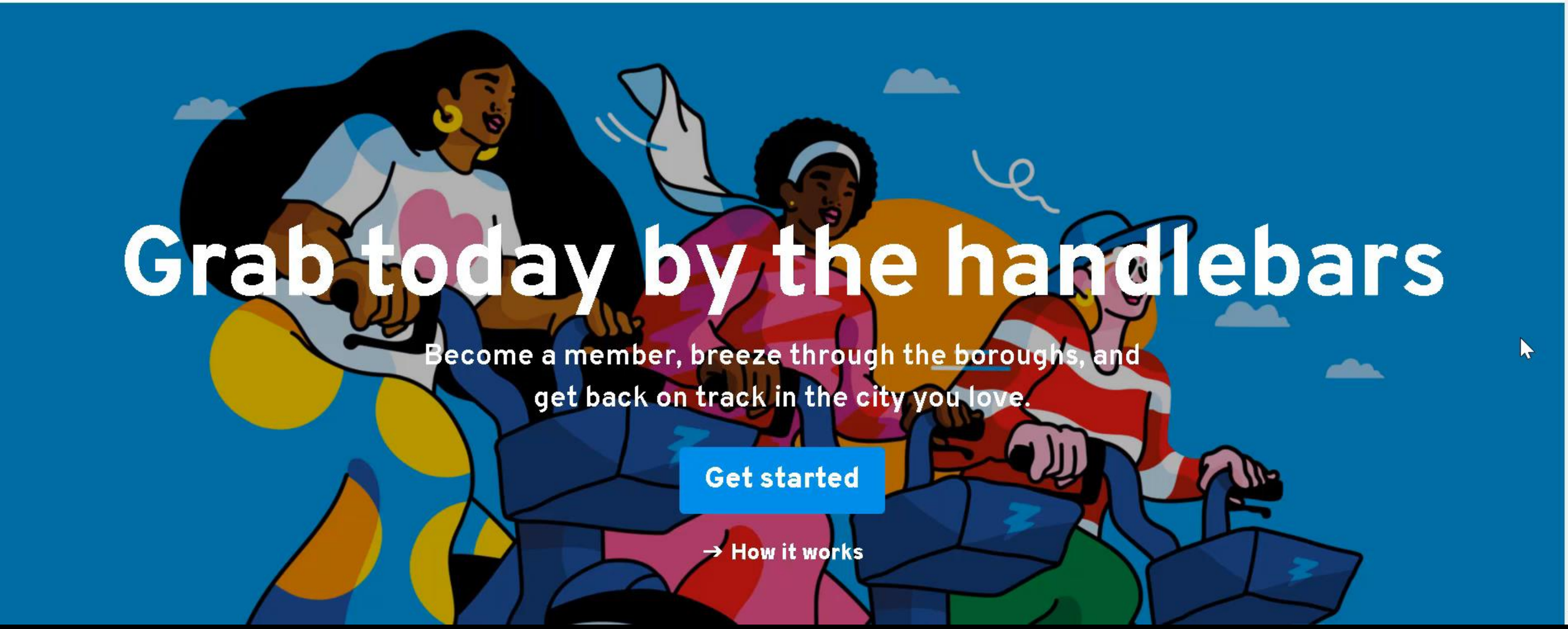
© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Isengard | Citi Bike: NYC's | CitiBike Analysis | S3 Management | AWS Glue DataB | AWS Glue Consc | Redshift

citibikenyc.com

citi bike® provided by lyft

How it Works | Pricing | Find a Bike | Explore NYC | Help | Log In | Join



Learn more with these sessions

Currently available

SESSION ID	SESSION TITLE
ANT318	The right tool for the job: Enabling analytics at scale at Intuit
ANT302	How Vyaire uses AWS analytics to scale ventilator production & save lives
ANT309	How Disney+ uses fast data ubiquity to improve the customer experience
ANT305	How Goldman Sachs uses an Amazon MSK backbone for its Transaction Banking Platform
ANT308	How FINRA operates PB-scale analytics on data lakes with Amazon Athena
ANT321	Top 5 best practices for data streaming with Amazon Kinesis
ANT303	Modernize log analytics with Amazon Elasticsearch Service
ANT312	Serverless data preparation with AWS Glue
ANT304	Migrating a legacy data warehouse to Amazon Redshift
ANT307	What's new with Amazon EMR
ANT203	What's new with Amazon Redshift
ANT320	What's new in Amazon Elasticsearch Service
ANT316	Security overview for Amazon Elasticsearch Service
ANT306	Run big data analytics faster at lower cost with Amazon EMR
ANT322	New use cases for Amazon Redshift

Thank you!



AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.



Please complete the session survey

AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.