Building Data Lake on AWS

Adir Sharabi

Solutions Architect, Amazon Web Services

SSID: Guest
Password: Cube@11999

Floor28 Agenda

Big Data Day 14 Oct	ML & DL Day 15 Oct	DevOps Day 16 Oct	DevOps Day 17 Oct	Databases Day 18 Oct
Technical Sessions	Technical Sessions	Technical Sessions	Technical Sessions	Technical Sessions
Serverless Data Workshop	SageMaker Workshop	K8s Workshop	Spot Workshop	
Big Data UG Meetup	ML&DL Meetup	DevOps Meetup		PyTorch Meetup

Builders Day Serverless backend 21 Oct	Builders Day AppSync, Alexa & IoT 22 Oct	Enterprise IT Day 23 Oct	GameDay 24 Oct
Technical Sessions	Technical Sessions	Technical Sessions	
Serverless Workshop	CDK Workshop		
Virtual assistants UG Meetup	AWS IL UG Meetup		

aws | FLOOR28 TEL-AVIV

Big Data Day Agenda

#	Time	Title	Speaker
1	9:30 - 10:15	Building Data Lake on AWS	Adir Sharabi
2	10:30 - 11:15	Store once, query thrice: Introduction to query engines on AWS	Daniel Haviv
3	11:30 – 12:15	Introduction to Real-Time Streaming Analytics - Amazon Kinesis State Of Union	Roy Ben Alta
4	12:30 - 13:15	From data to insights	Orit Alul
5	15:00 – 18:00	Serverless Data Processing Workshop	Adir Sharabi
6	18:00 – 20:00	Big Data User Group Meetup	

Your Data Sources

Multiple sources and formats… and growing everyday

Documents and files


Clickstream data


Mobile app data


Spreadsheets


Infrastructure logs


Social media data

Records


Amazon RDS


ERP


Amazon DynamoDB


On Premises databases

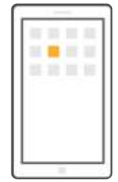

Amazon Redshift

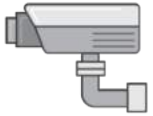
Streams


AWS IoT


WEB Clickstream

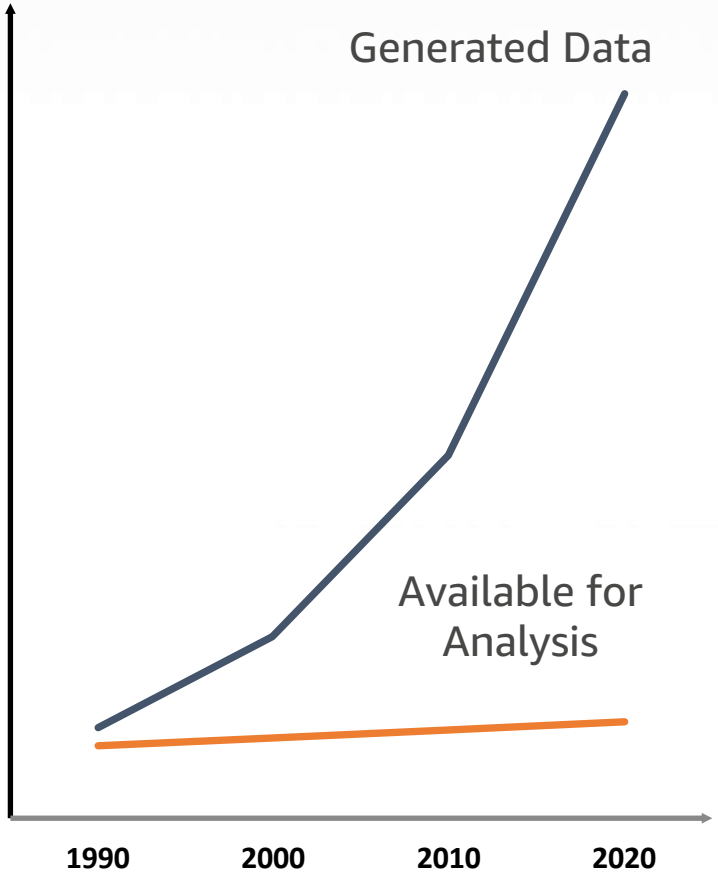

Device data


Mobile Apps


Sensor data

Data Challenges

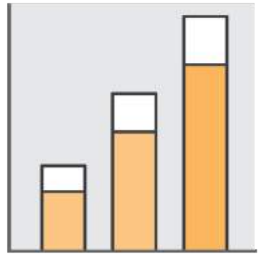
Data Visibility



Multiple consumers and requirements

Data Scientists

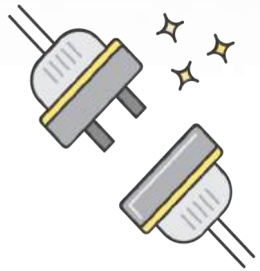
Business Users



Analysts

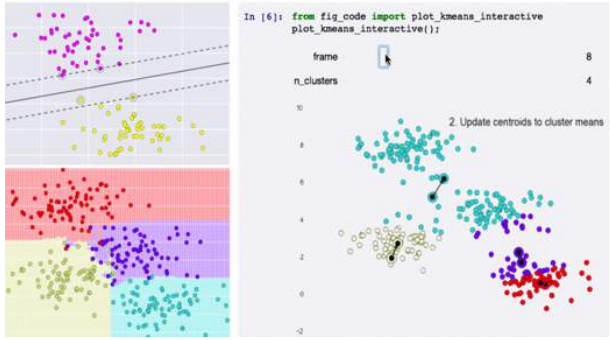
Applications

Multiple Access Mechanisms



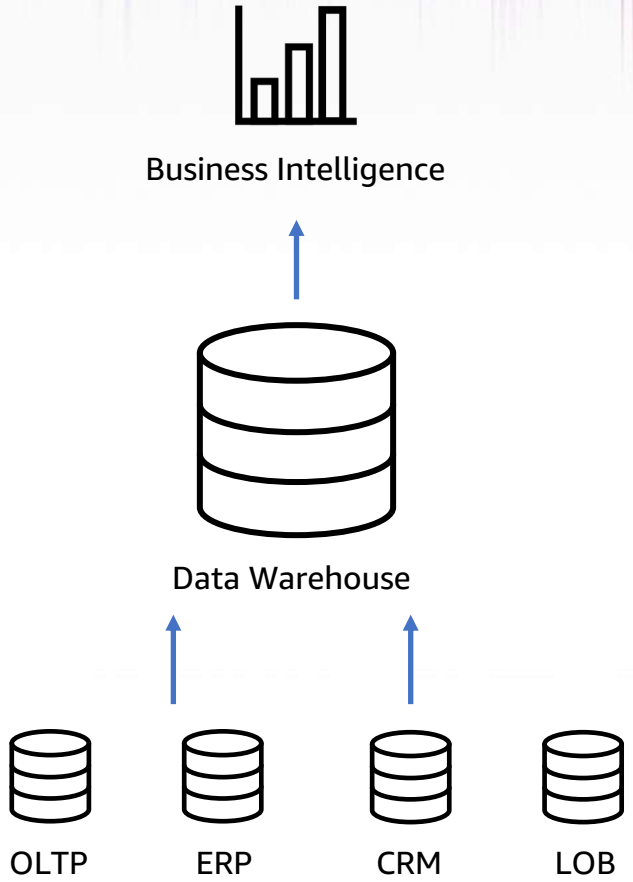
API Access

BI Tools



Notebooks

Traditionally, Analytics Used to Look Like This



Relational data

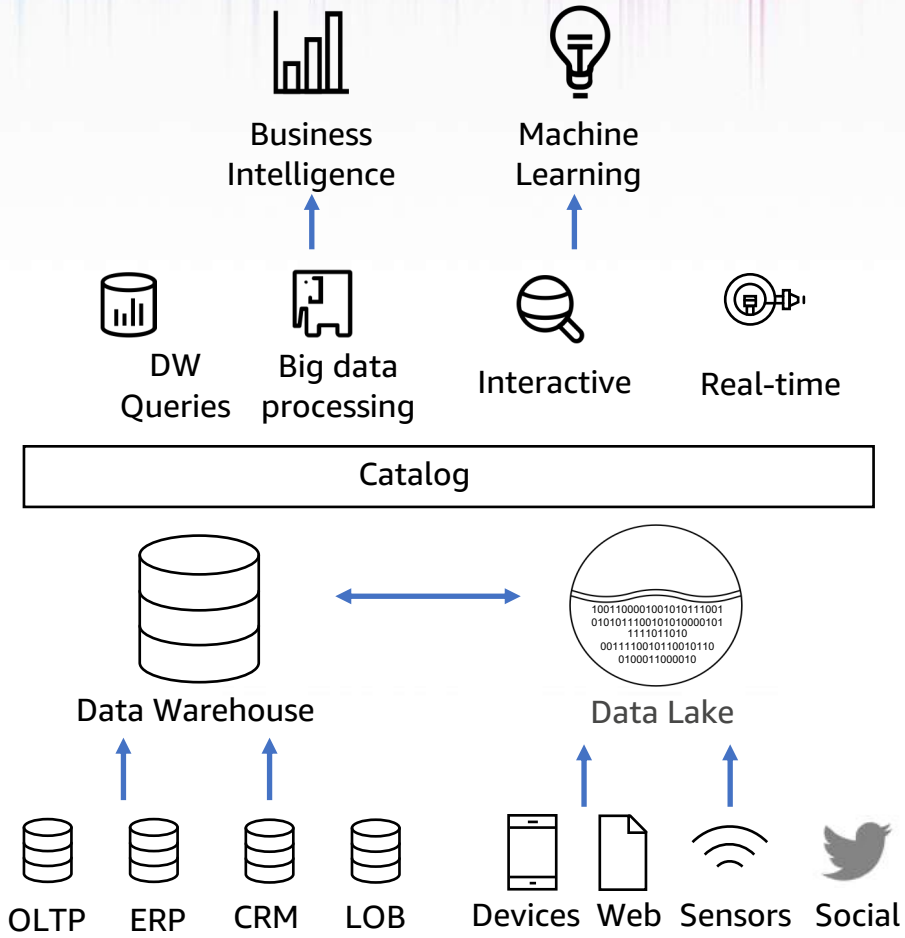
Schema defined prior to data load

TBs-PBs Scale

Operational reporting and ad hoc

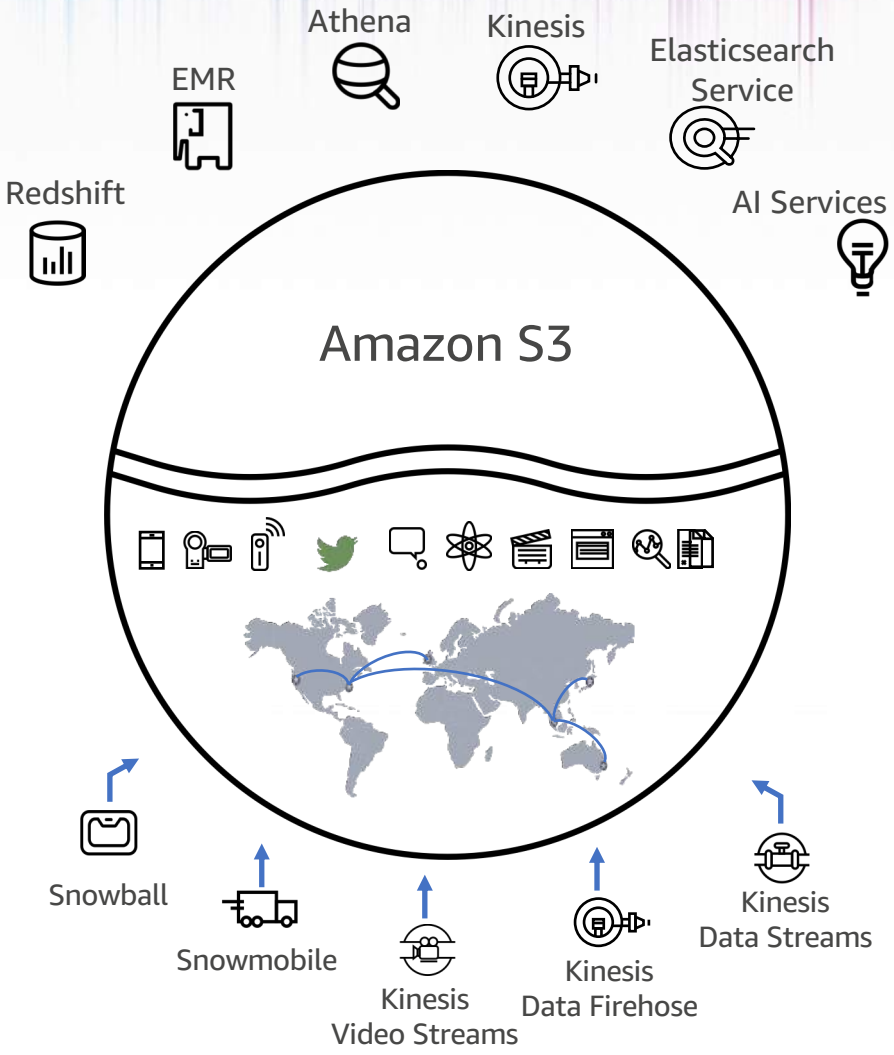
Large initial capex + $10K–$50K/TB/Year

Data Lakes Extend the Traditional Approach



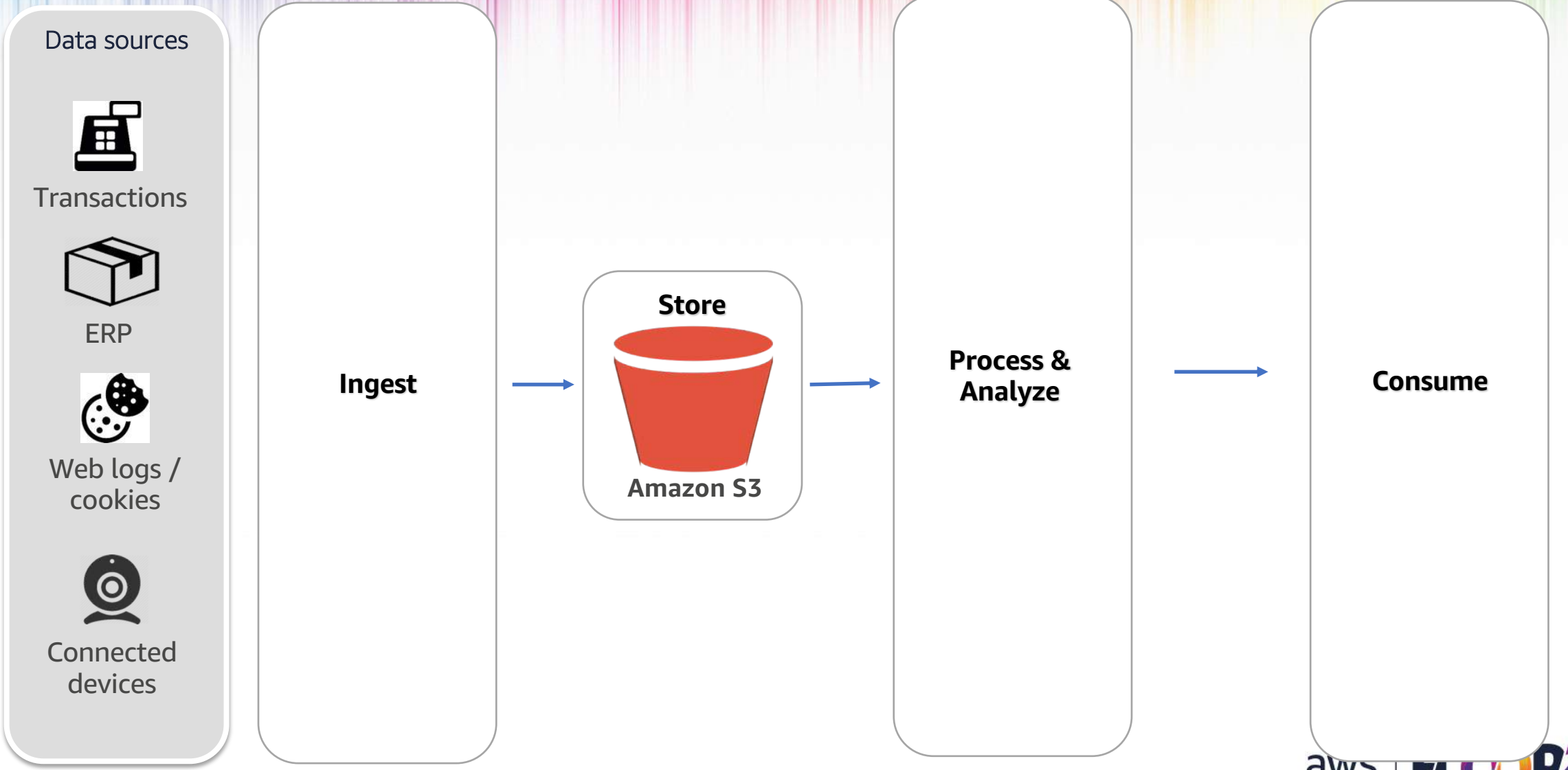
- Relational and non-relational data
- Schema defined during analysis
- Scale storage and compute independently
- Diverse analytical engines to gain insights
- Designed for low-cost storage and analytics

Amazon S3 as Data Lakes Storage Layer

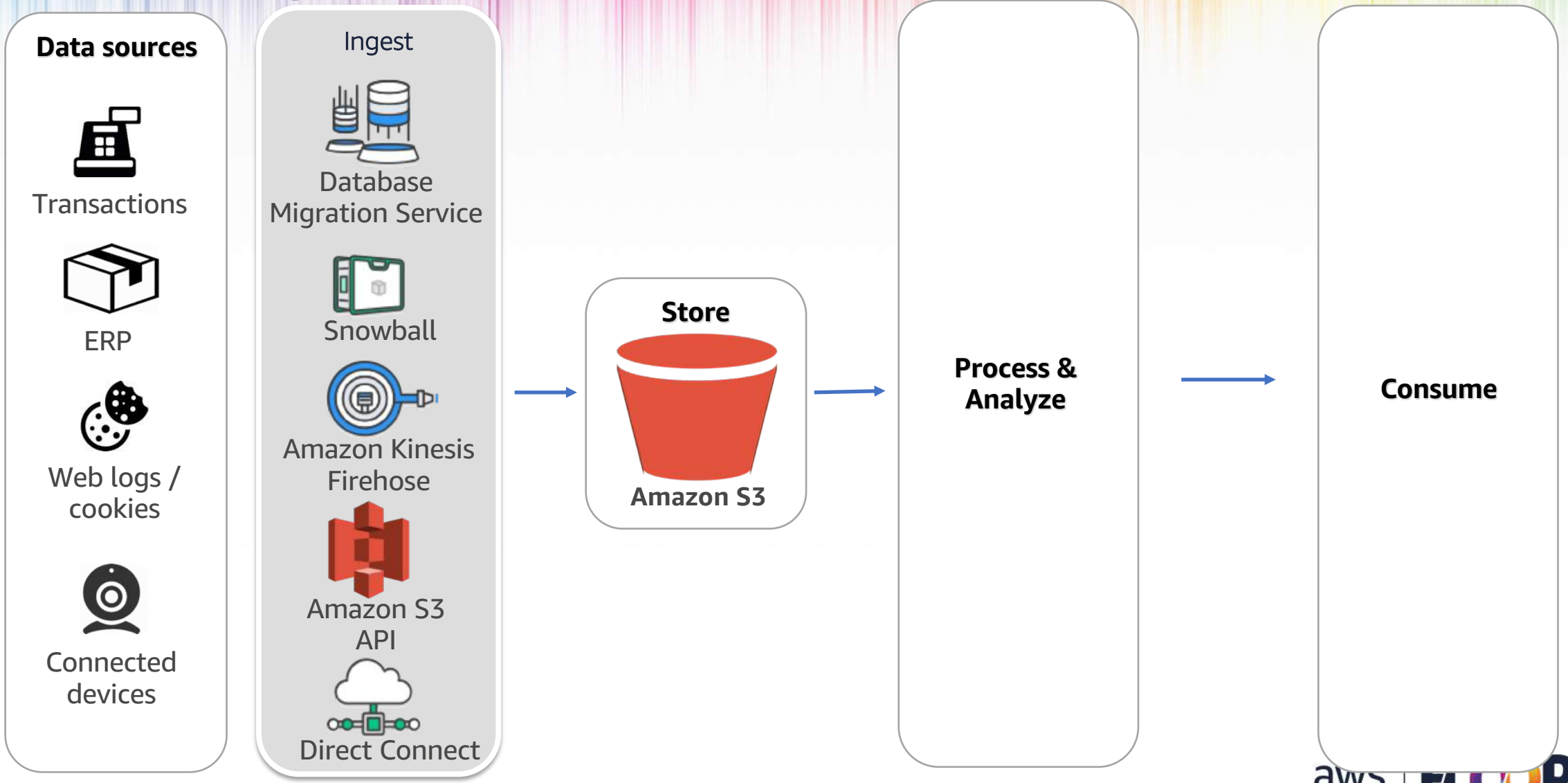


- Many ways to bring all kinds of data
- Unmatched durability and availability at EB scale
- Best security, compliance, and audit capabilities
- Integration with Big Data Tools
- Run any analytics on the same data without movement
- Cost effective - Store data at $0.023 / GB / Month

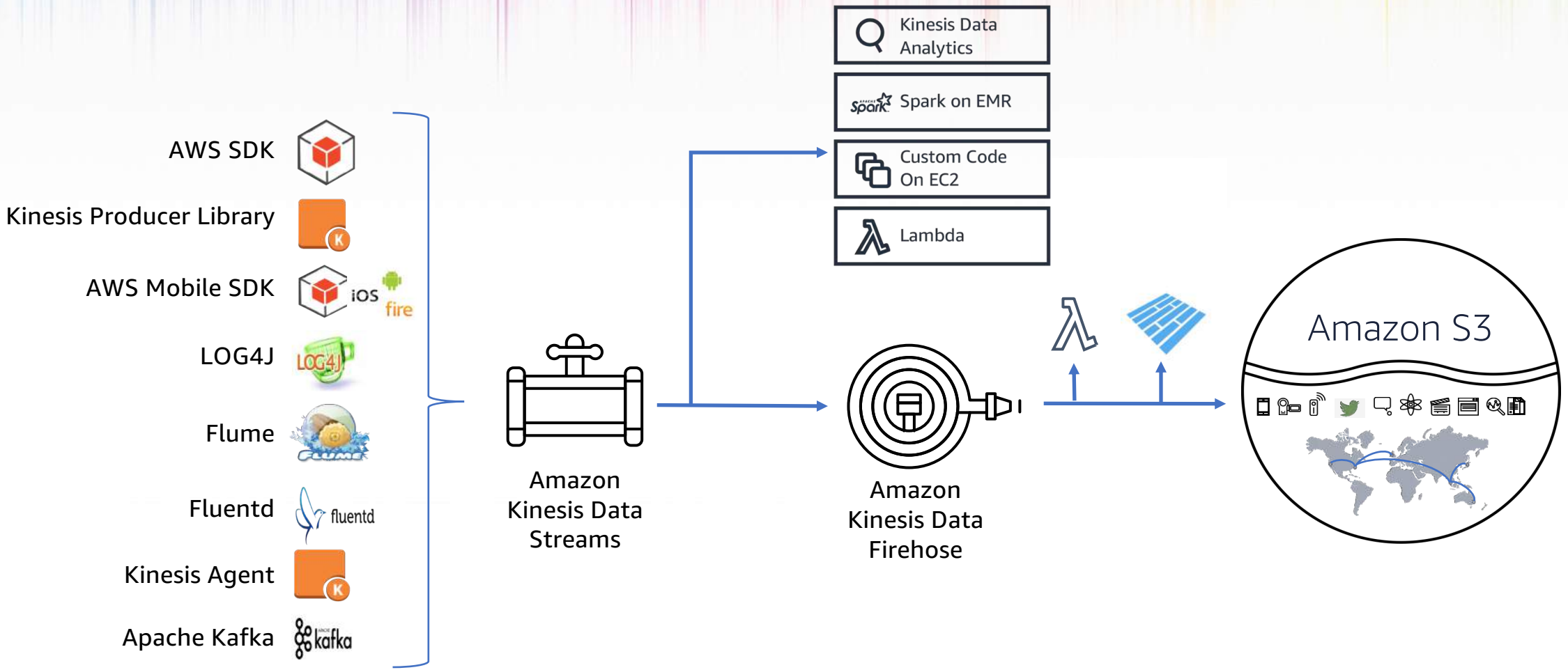
Simplified Big Data Pipeline



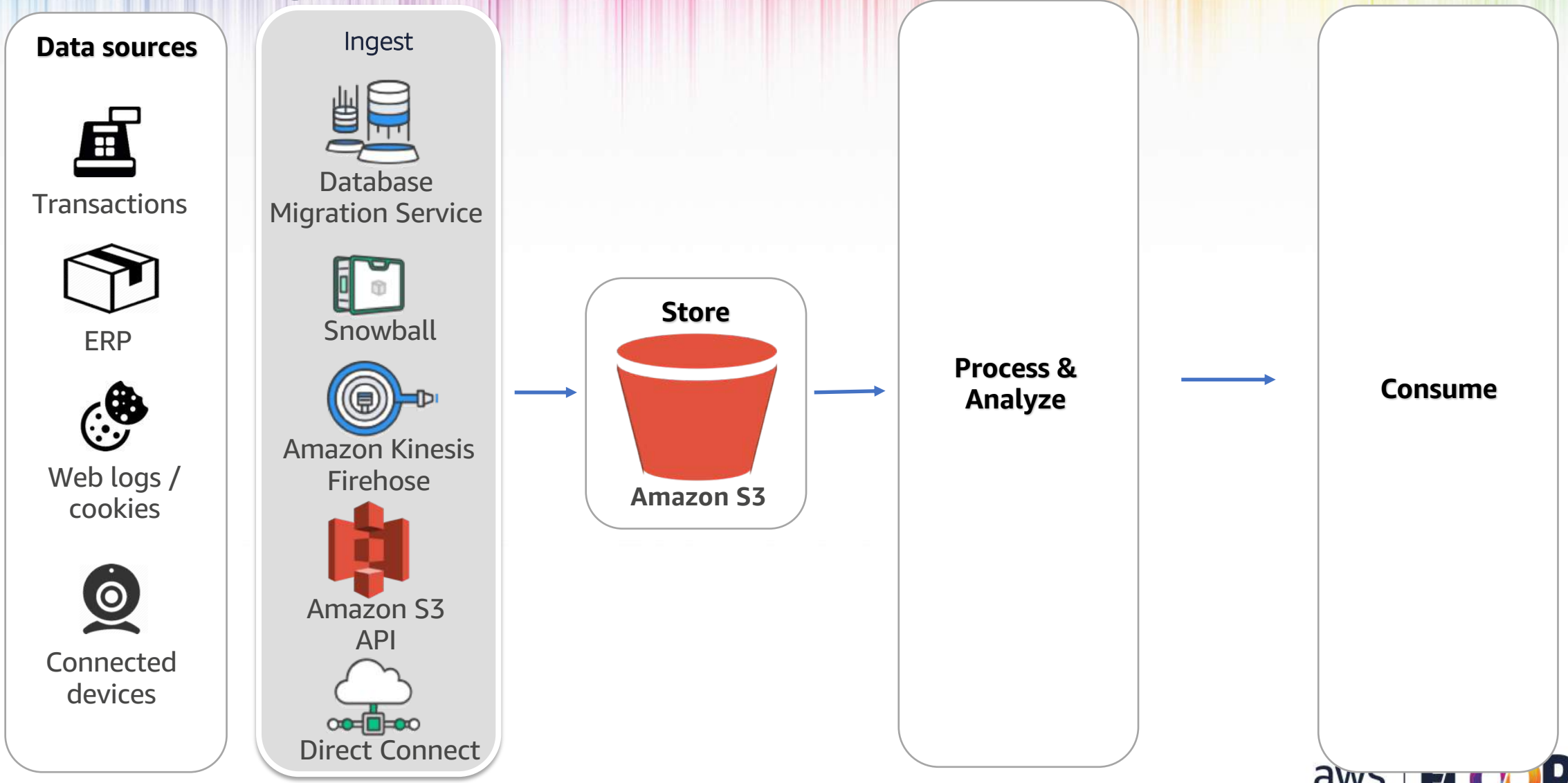
Lots of ingestion tools



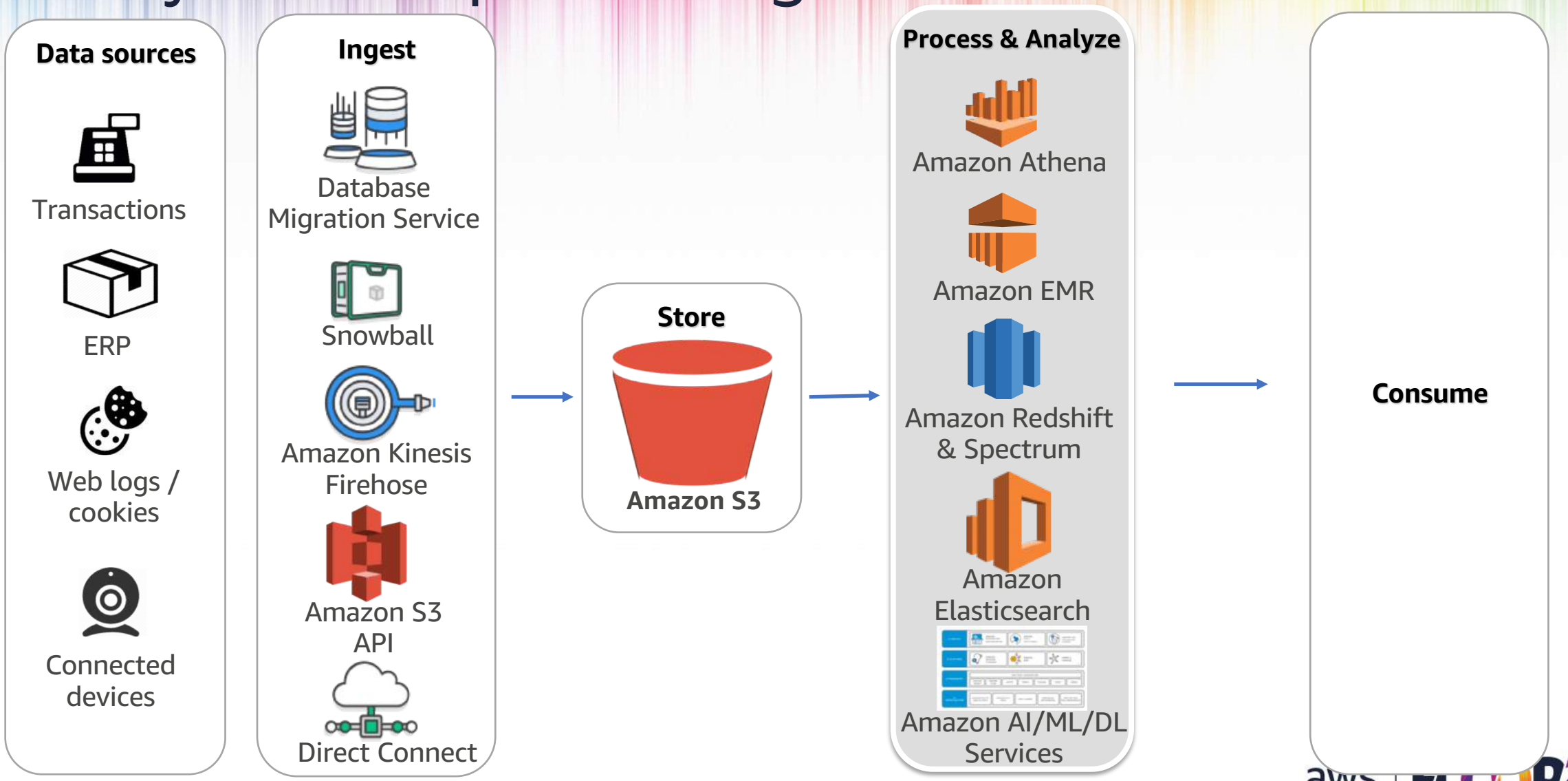
Real-time data movement and Data Lakes on AWS



Lots of ingestion tools



Variety of data processing tools



Amazon Athena – interactive analysis

Interactive query service to analyze data in Amazon S3 using standard SQL

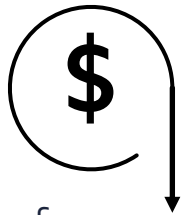
No infrastructure to set up or manage and no data to load

Ability to run SQL queries on data archived in Amazon Glacier (coming soon)

Query instantly

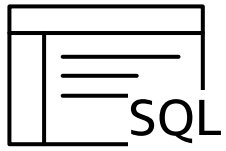
Zero setup cost; just point to Amazon S3 and start querying

Pay per query



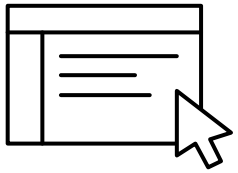
Pay only for queries run; save 30%–90% on per-query costs through compression

Open



ANSI SQL interface, JDBC/ODBC drivers, multiple formats, compression types, and complex joins and data types

Easy



Serverless: zero infrastructure, zero administration
Integrated with Amazon QuickSight

Amazon EMR – big data processing

Analytics and ML at scale

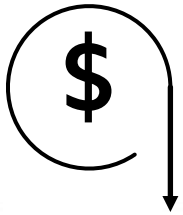
19 open-source projects: Apache Hadoop, Spark, HBase, Presto, and more

Enterprise-grade security

Latest versions

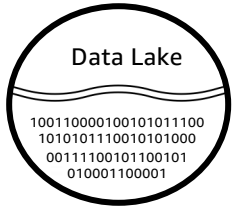
Updated with the latest open source frameworks within 30 days of release

Low cost



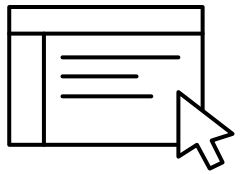
Flexible billing with per-second billing, Amazon EC2 Spot, Reserved Instances, and Auto Scaling to reduce costs 50%-80%

Use Amazon S3 storage



Process data directly in the Amazon S3 data lake securely with high performance using the EMRFS connector

Easy



Launch fully managed Hadoop & Spark in minutes; no cluster setup, node provisioning, cluster tuning

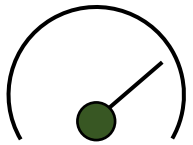
aws | FLOOR28 TEL-AVIV

Amazon Redshift – data warehousing

Fast, powerful, simple, and fully managed data warehouse at 1/10 the cost

Massively parallel, scale from gigabytes to petabytes

Fast at scale



Columnar storage technology to improve I/O efficiency and scale query performance

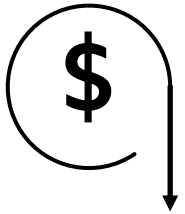
Open file formats

Analyze optimized data formats on the latest SSD, and all open data formats in Amazon S3

Secure

Audit everything; encrypt data end-to-end; extensive certification and compliance

Inexpensive

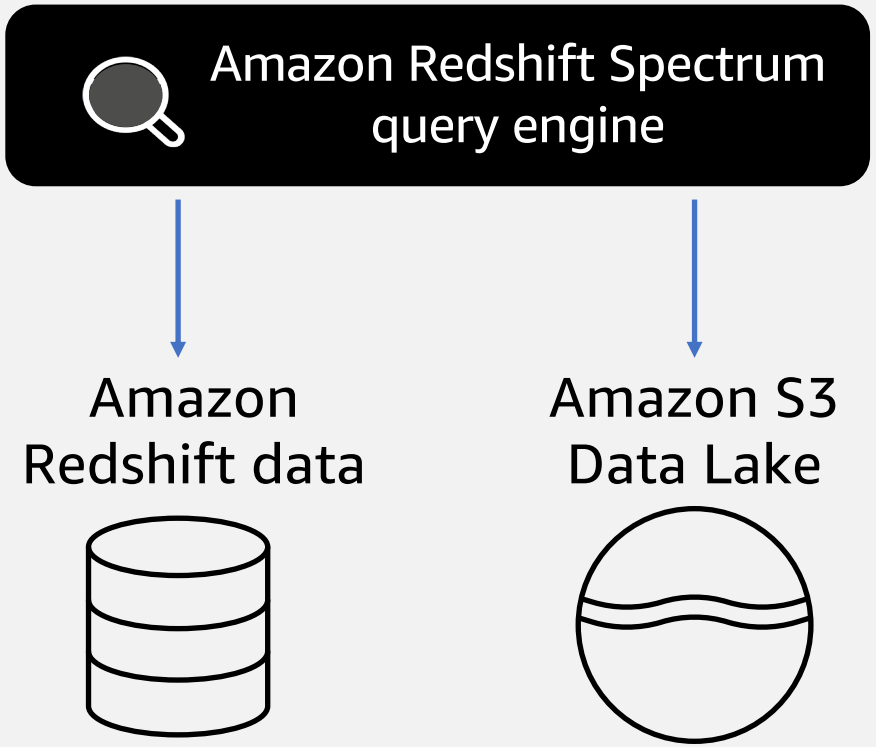


As low as $1,000 per terabyte per year, 1/10 the cost of traditional data warehouse solutions; start at $0.25 per hour

aws | FLOOR28 TEL-AVIV

Amazon Redshift Spectrum

Extend the data warehouse to exabytes of data in Amazon S3 data lake



- Exabyte Redshift SQL queries against Amazon S3
- Join data across Redshift and Amazon S3
- Scale compute and storage separately
- Stable query performance and unlimited concurrency
- CSV, ORC, Grok, Avro, & Parquet data formats
- Pay only for the amount of data scanned

JDBC / ODBC

Amazon Redshift

Hot data

Query directly on data lake

COPY commands

1 2 3 4 ... N

Redshift Spectrum

Scale-out serverless compute

Multiple ways to consume the data

Data sources

- Transactions
- ERP
- Web logs / cookies
- Connected devices

Ingest

- Database Migration Service
- Snowball
- Amazon Kinesis Firehose
- Amazon S3 API
- Direct Connect

Store

- Amazon S3

Process & Analyze

- Amazon Athena
- Amazon EMR
- Amazon Redshift & Spectrum
- Amazon Elasticsearch
- Amazon AI/ML/DL Services

Consume

- Jupyter, Zeppelin, HUE
- Amazon API Gateway
- Amazon QuickSight
- BI Tools

Because data is NEVER perfect

Clean

Transform

Concatenate

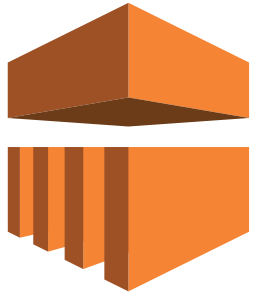
Convert to better formats

Schedule transformations

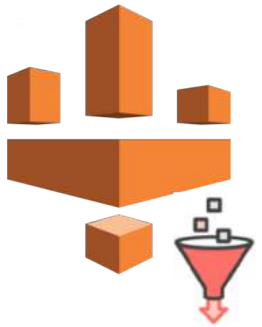
Event-driven transformations

Transformations expressed as code

AWS Lambda
Trigger-based Code Execution



Amazon EMR
Spark and Hive running on EMR

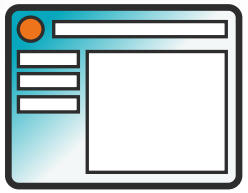


AWS Glue
Event based Server-less ETL engine

AWS Glue

Data Catalog

Job Authoring



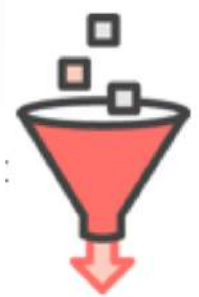
Develop

Auto-generates ETL code

Python/Scala and Apache Spark

Edit, debug, and share

Job Execution



Deploy

Serverless execution

Flexible scheduling

Monitoring and alerting

ETL when you need it

Data sources

- Transactions
- ERP
- Web logs / cookies
- Connected devices

Ingest

- Database Migration Service
- Snowball
- Amazon Kinesis Firehose
- Amazon S3 API
- Direct Connect

Store

- Amazon S3

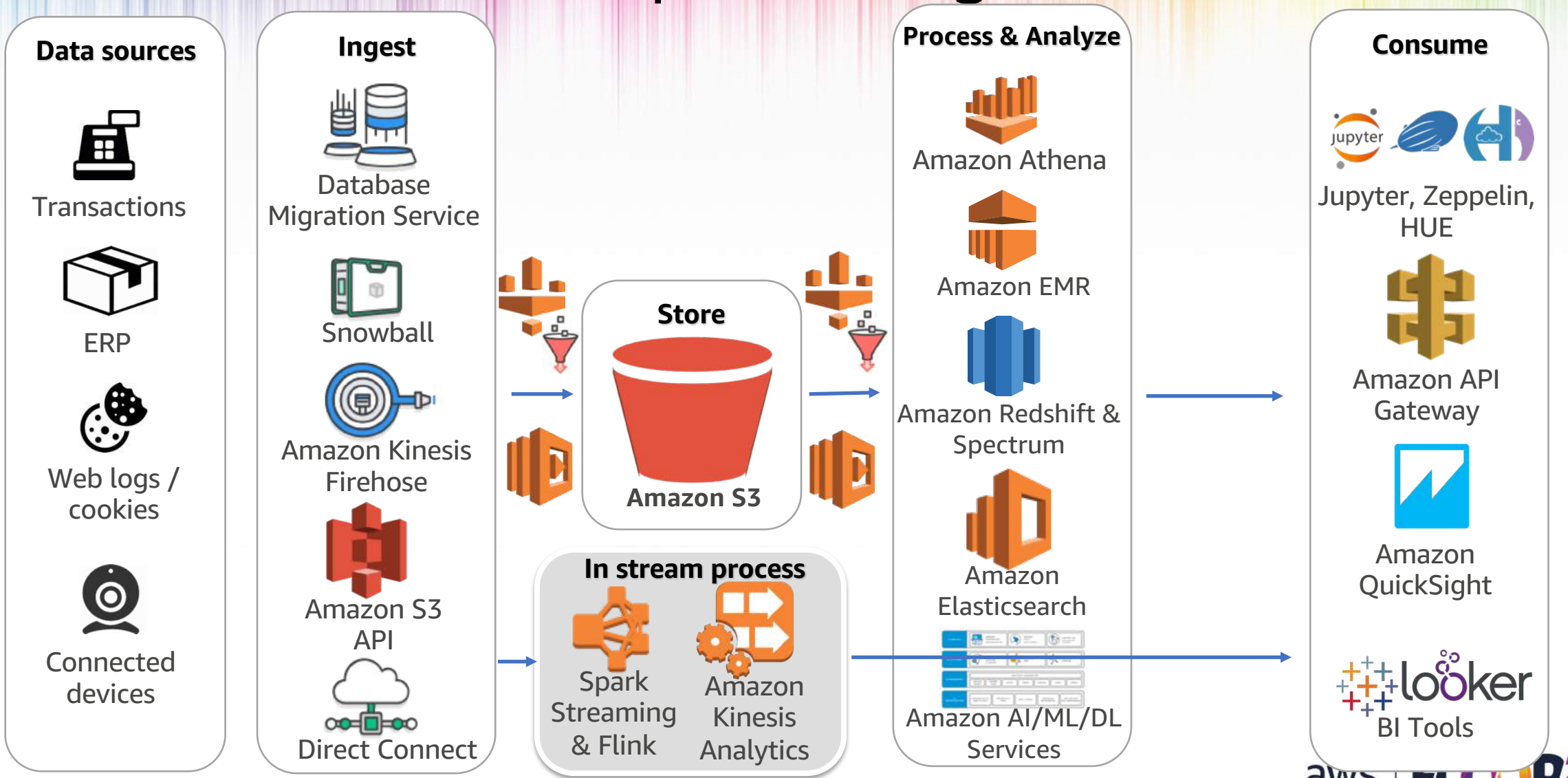
Process & Analyze

- Amazon Athena
- Amazon EMR
- Amazon Redshift & Spectrum
- Amazon Elasticsearch
- Amazon AI/ML/DL Services

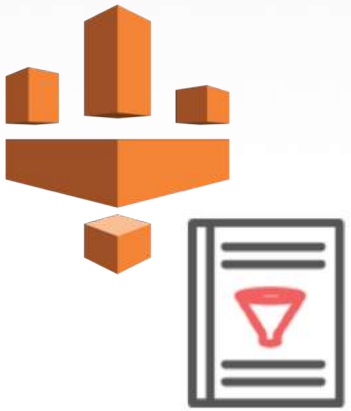
Consume

- Jupyter, Zeppelin, HUE
- Amazon API Gateway
- Amazon QuickSight
- BI Tools

Realtime - in-stream processing



AWS Glue Data Catalog



Central Metadata Catalog for the data lake

One per account

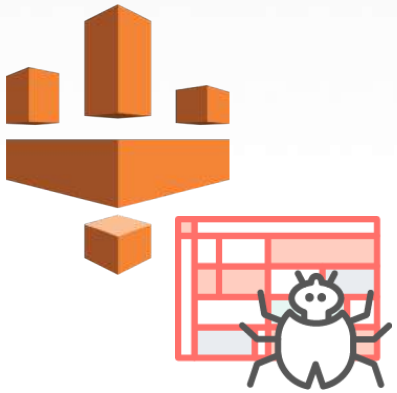
Allows you to share metadata between Amazon Athena, Amazon Redshift Spectrum, EMR & JDBC sources

Serverless

We added a few extensions:

- **Search** over metadata for data discovery
- **Manage Connections** – JDBC URLs, credentials
- **Classification** for identifying and parsing files
- **Versioning** of table metadata as schemas evolve and other metadata are updated

AWS Glue Data Catalog Crawlers



Catalogs Your Data

Crawlers automatically build your Data Catalog and keep it in sync

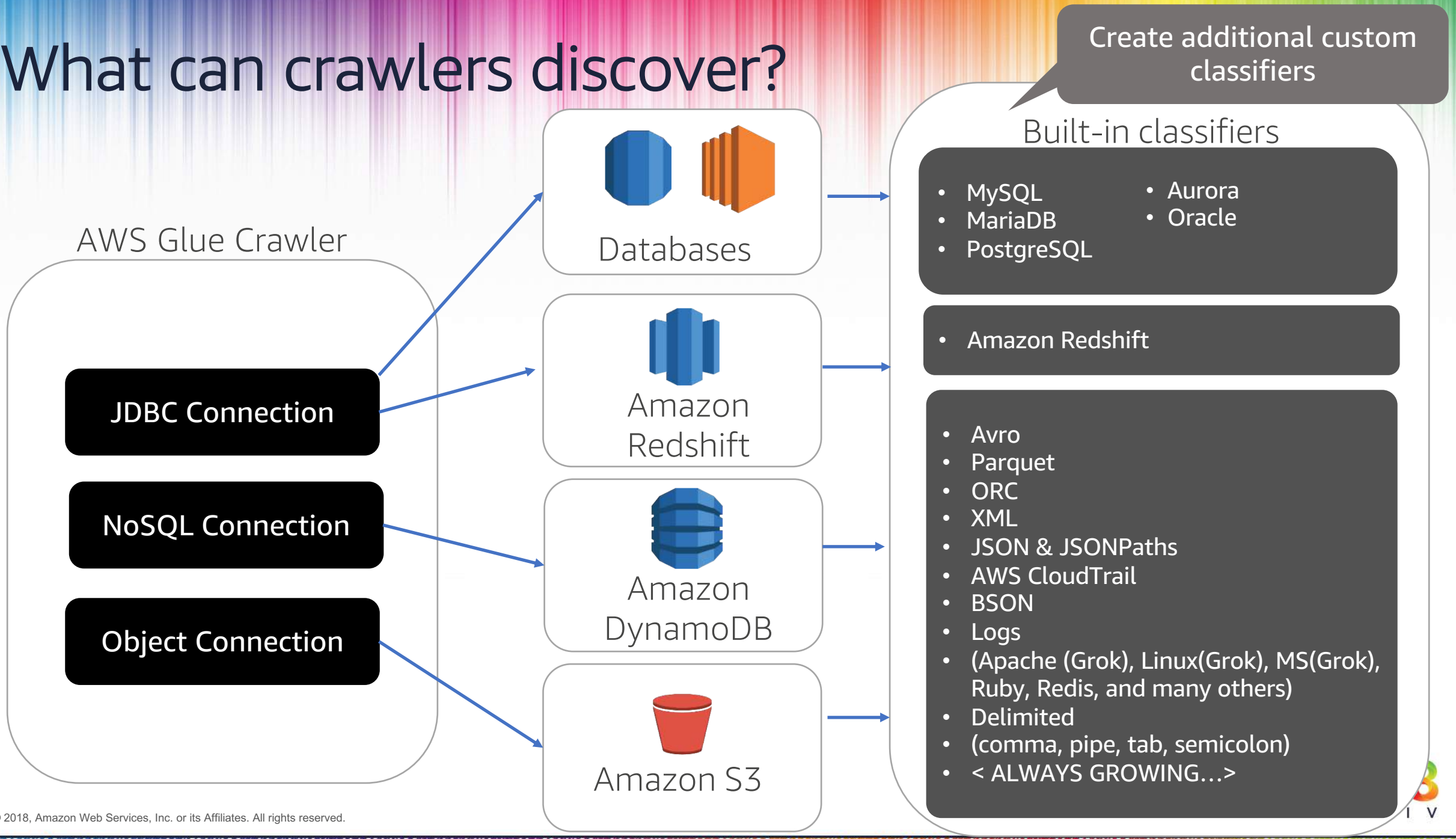
Automatically discover new data, extracts schema definitions

- Detect schema changes and version tables
- Detect Hive style partitions on Amazon S3

Built-in classifiers for popular types; custom classifiers using Grok expression

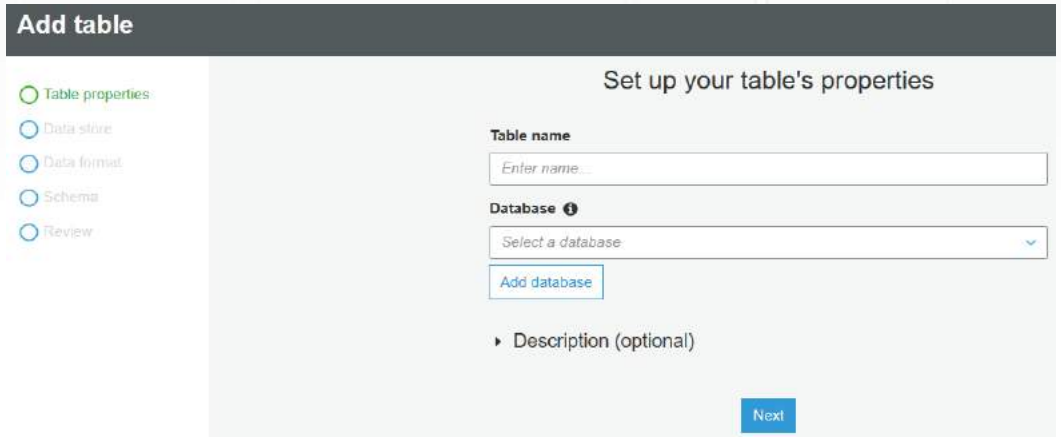
Run ad hoc or on a schedule; serverless – only pay when crawler runs

What can crawlers discover?



Other ways of populating the catalog

Create table manually

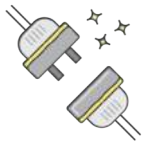


DDL statement (in Amazon Athena or Amazon EMR)

```
CREATE EXTERNAL TABLE IF NOT EXISTS elb_logs_raw_native_part (
  request_timestamp string,
  elb_name string,
  request_ip string,
  request_port int,
  backend_ip string,
  backend_port int,
  request_processing_time double,
  backend_processing_time double,
  client_response_time double,
  elb_response_code string,
  backend_response_code string,
  received_bytes bigint,
  sent_bytes bigint,
  request_verb string,
  url string,
  protocol string,
  user_agent string,
  ssl_cipher string,
  ssl_protocol string )
PARTITIONED BY(year string, month string, day string)
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.RegexSerDe'
WITH SERDEPROPERTIES (
        'serialization.format' = '1','input.regex' = '([^ ]*) ([^ ]*) ([^ ]*):([0-9]*) ([^ ]*):([0-9]*) ([.0-9]*) ([.0-
LOCATION 's3://athena-examples/elb/raw/';
```

Use Ctrl + Enter to run query, Ctrl + Space to autocomplete

Run Query | Save As | Format Query | New Query (Run time: 2.03 seconds, Data scanned: 0KB)



Call the AWS Glue CreateTable API

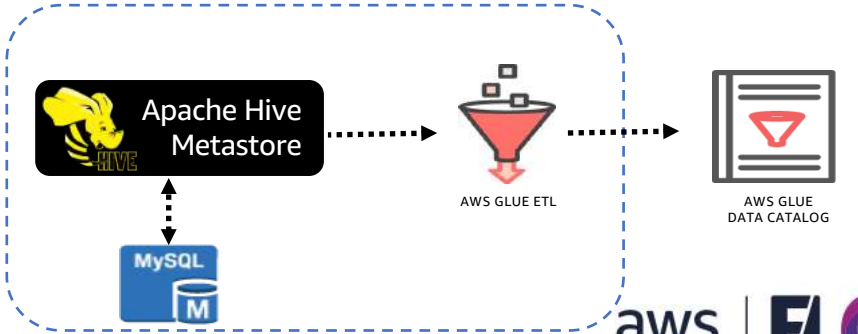
CreateTable

Creates a new table definition in the Data Catalog.

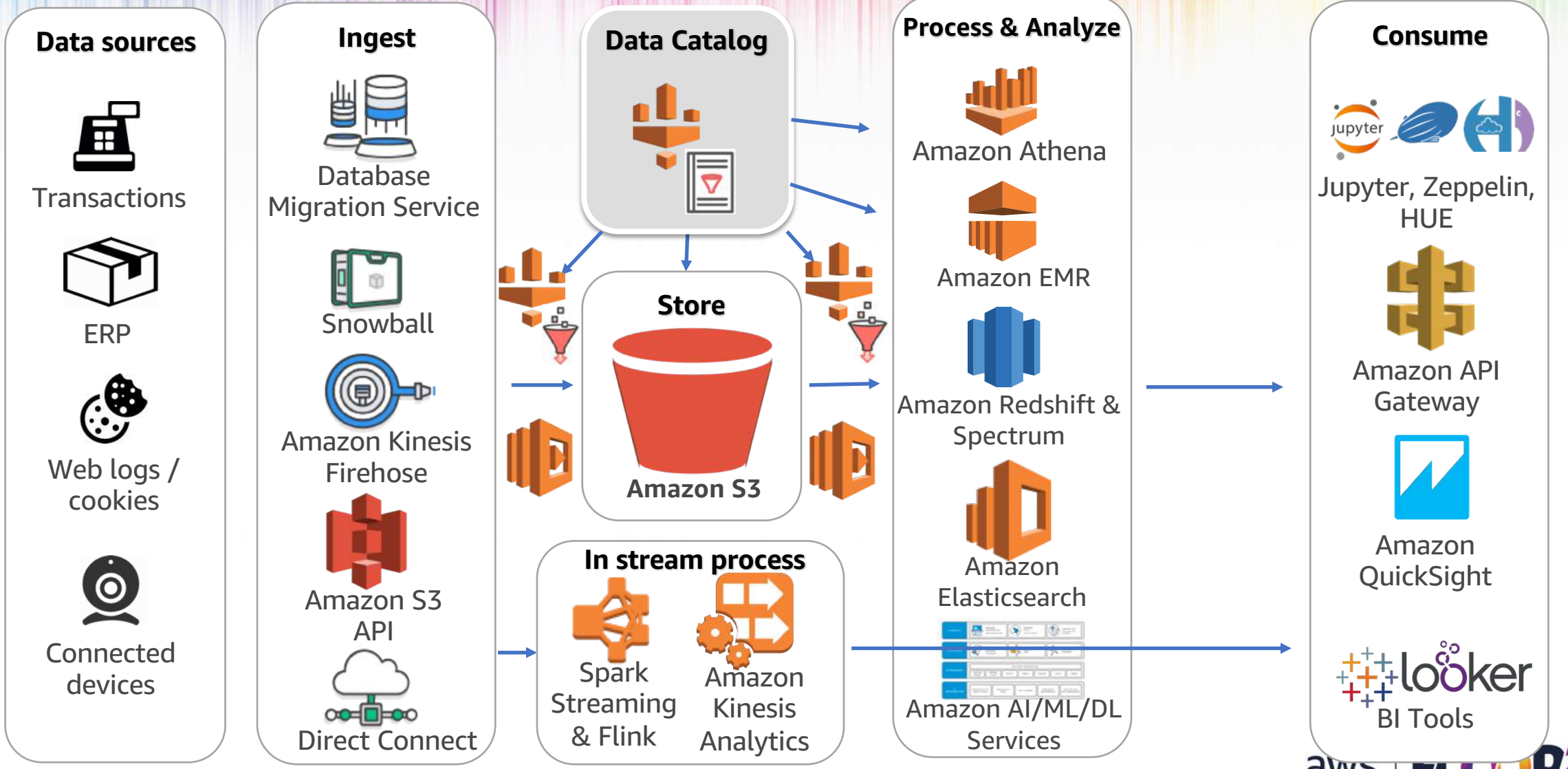
Request Syntax

```
{
   "CatalogId": "string",
   "DatabaseName": "string",
   "TableInput": {
      "Description": "string",
```

Import from Apache Hive Metastore



Write once, catalog once, read multiple, ETL Anywhere



Core Tenets

- Data lakes and data warehouses complement each other
- Loose Coupling, but highly performant
    - Storage, analytics, metadata management, etc..
- Choosing the best tool for the job
- Future-proof your analytics
- Elasticity and multiple clusters for dedicated purposes
- Replace capacity planning with a consumption model
- Don’t forget metadata management

Thank You!

Adir Sharabi

aws | FLOOR28 TEL-AVIV

GAME DAY

PUT YOUR SKILLS TO THE TEST

OCT 24

Register now: bit.ly/Floor28GameDay

SSID: Guest
Password: Cube@11999