Deep Dive on Amazon Redshift

Agenda

Deep inside Redshift Architecture	10%
Performance tuning	5%
Integration with AWS data services	5%
Redshift Spectrum	10%
Redshift Echo System	15%
Redshift at Manchester Airport Group	45%
Summary + Q&A	10%

Redshift Architecture

Fast

Cost Efficient

Simple

Elastic

Secure

Compatible

Amazon Redshift

Fast, simple, cost-effective data warehousing.

Managed Massively Parallel Petabyte Scale Data Warehouse

Streaming Backup/Restore to S3

Load data from S3, DynamoDB and EMR

Extensive Security Features

Scale from 160 GB -> 2 PB Online

Amazon Redshift Cluster Architecture

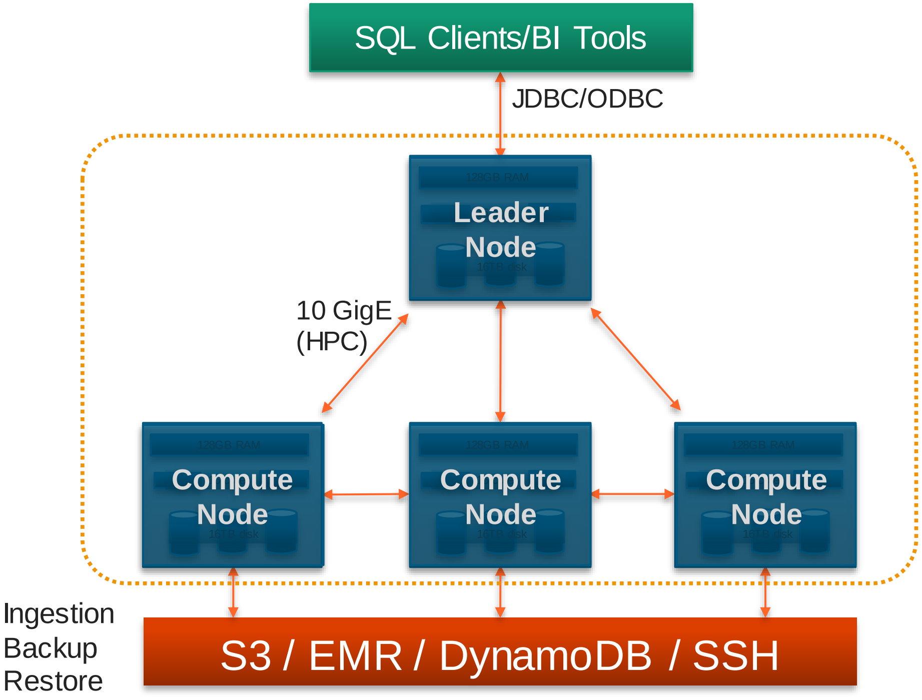
Massively parallel, shared nothing

Leader node

- SQL endpoint
- Stores metadata
- Coordinates parallel SQL processing

Compute nodes

- Local, columnar storage
- Executes queries in parallel
- Load, backup, restore
- 2, 16 or 32 slices



Design for Queryability

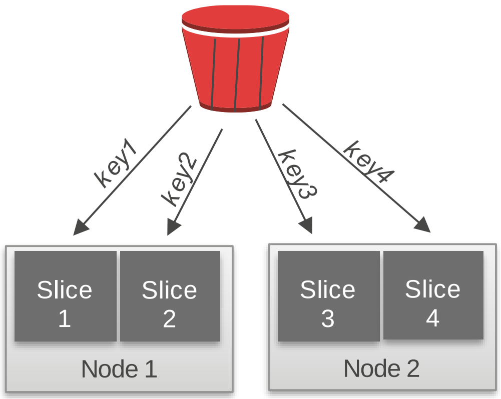
- *Equally* on *each slice*
- *Minimum* amount of work
- Use *just enough* cluster resources

Do an *Equal* Amount of Work on Each Slice

Choose Best Table Distribution Style

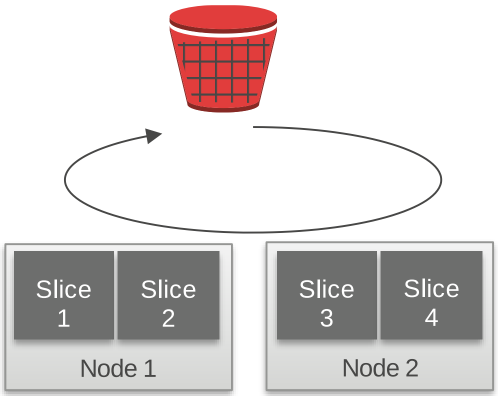
Key

Same key to same location



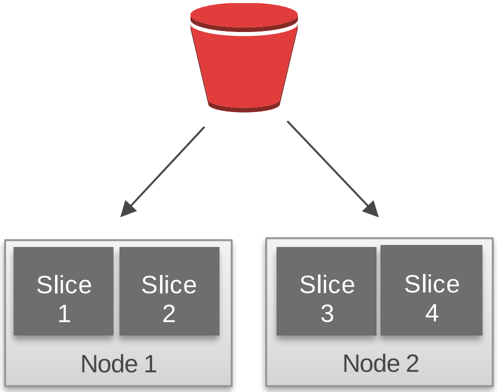
Even

Round robin distribution



All

All data on every node



Do the *Minimum* Amount of Work on Each Slice

Reduced I/O = Enhanced Performance

Columnar storage

+

Large data block sizes

+

Data compression

+

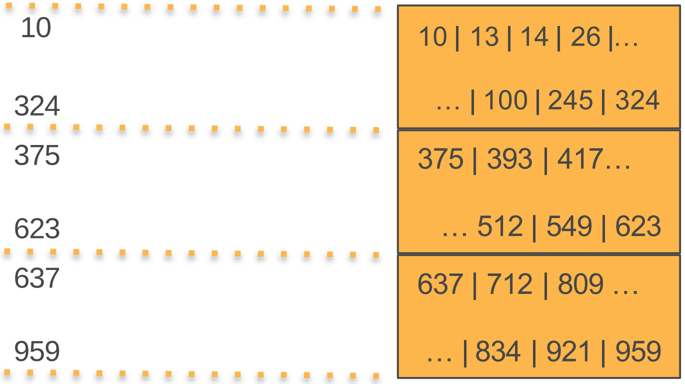
Zone maps

+

Direct-attached storage

```
analyze compression listing;

  Table   |      Column      | Encoding
----------+------------------+----------
 listing  | listid           | delta
 listing  | sellerid         | delta32k
 listing  | eventid          | delta32k
 listing  | dateid           | bytedict
 listing  | numtickets       | bytedict
 listing  | priceperticket   | delta32k
 listing  | totalprice       | mostly32
 listing  | listtime         | raw
```



10

324

375

623

637

959

10 | 13 | 14 | 26 |…

… | 100 | 245 | 324

375 | 393 | 417…

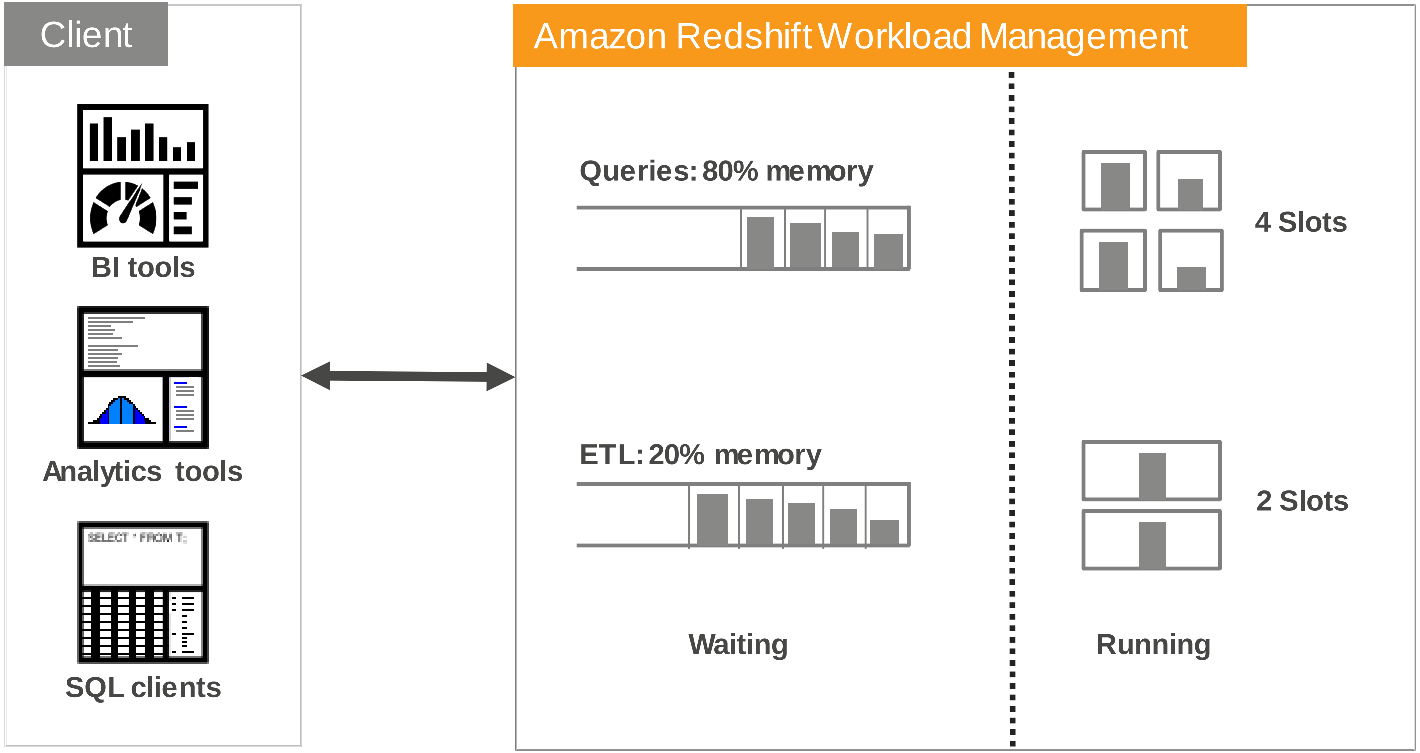
… 512 | 549 | 623

637 | 712 | 809 …

… | 834 | 921 | 959

Use Cluster Resources Efficiently to Complete as Quickly as Possible

Workload Management



Client

BI tools

Analytics tools

SQL clients

Amazon Redshift Workload Management

Queries: 80% memory

4 Slots

ETL: 20% memory

2 Slots

Waiting

Running

80/4 = **20% per slot**

20/2 = **10% per slot**

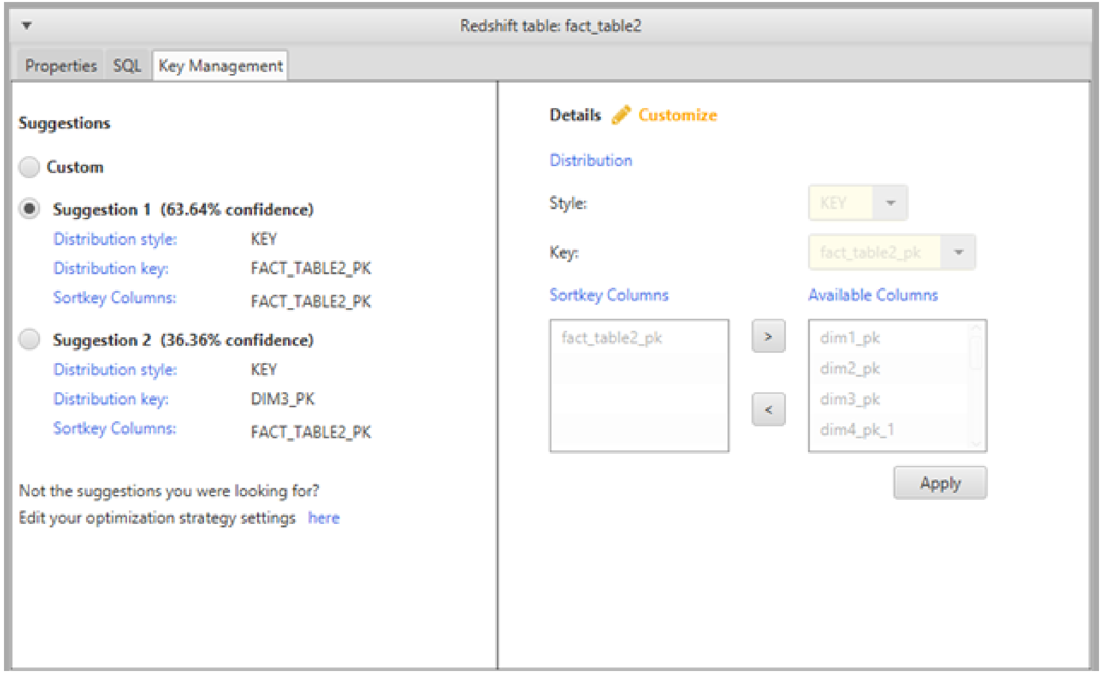
Redshift Performance Tuning

Redshift Playbook

Part 1: Preamble, Prerequisites, and **Prioritization**
Part 2: **Distribution** Styles and Distribution Keys
Part 3: Compound and Interleaved **Sort Keys**
Part 4: Compression **Encodings**
Part 5: Table Data **Durability**

amzn.to/2quChdM

Optimizing Amazon Redshift by Using the AWS Schema Conversion Tool



Suggestions

○ Custom

● **Suggestion 1 (63.64% confidence)**

Distribution style:	KEY
Distribution key:	FACT_TABLE2_PK
Sortkey Columns:	FACT_TABLE2_PK

○ **Suggestion 2 (36.36% confidence)**

Distribution style:	KEY
Distribution key:	DIM3_PK
Sortkey Columns:	FACT_TABLE2_PK

Not the suggestions you were looking for?
Edit your optimization strategy settings here

Details Customize

Distribution

Style: KEY

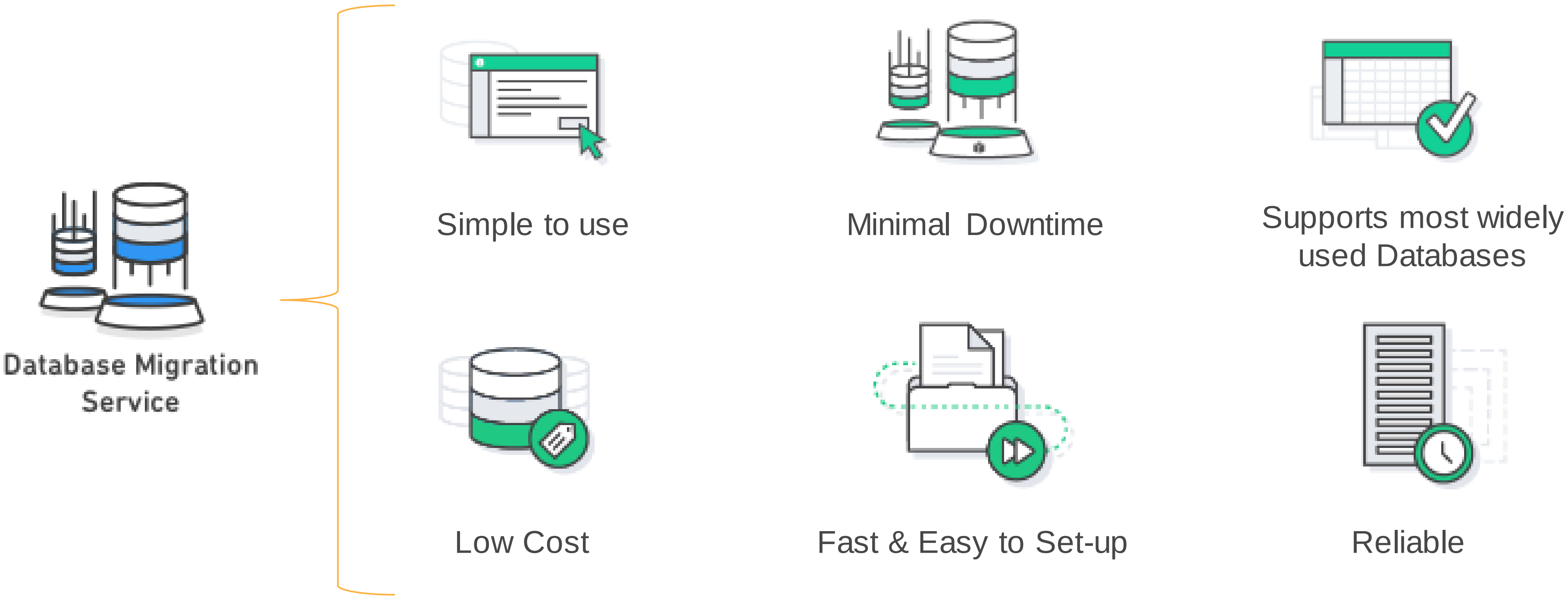
Key: fact_table2_pk

Sortkey Columns	Available Columns
fact_table2_pk	dim1_pk dim2_pk dim3_pk dim4_pk_1

amzn.to/2sTYow1

Ingestion, ETL & BI

Getting data to Redshift using AWS DMS



Database Migration Service

Simple to use

Minimal Downtime

Supports most widely used Databases

Low Cost

Fast & Easy to Set-up

Reliable

Loading data from S3

- Splitting Your Data into Multiple Files
- Uploading Files to Amazon S3
- Using the COPY Command to Load from Amazon S3

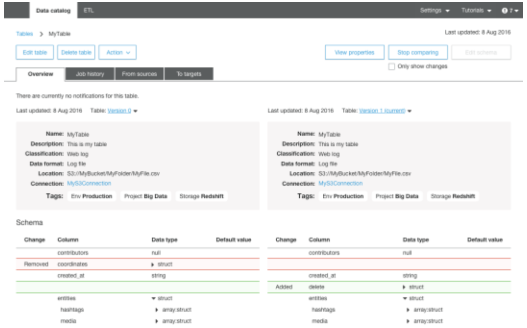
ETL on Redshift

AWS Glue

Fully-managed data catalog and ETL service

Integrated with:

S3, RDS, Redshift & any
JDBC-compliant data store



Last updated: 8 Aug 2016

□ Only show changes

There are currently no notifications for this table.

Last updated: 8 Aug 2016 Table: Version 0 ▾

Name: MyTable
Description: This is my table
Classification: Web log
Data format: Log file
Location: S3://MyBucket/MyFolder/MyFile.csv
Connection: MyS3Connection
Tags: Env **Production** Project **Big Data** Storage **Redshift**

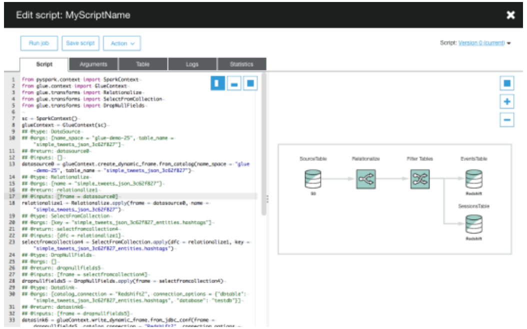
Last updated: 8 Aug 2016 Table: Version 1 (current) ▾

Name: MyTable
Description: This is my table
Classification: Web log
Data format: Log file
Location: S3://MyBucket/MyFolder/MyFile.csv
Connection: MyS3Connection
Tags: Env **Production** Project **Big Data** Storage **Redshift**

Schema

Change	Column	Data type	Default value
	contributors	null	
Removed	coordinates	▸ struct	
	created_at	string	
	entities	▾ struct	
	hashtags	▸ array:struct	
	media	▸ array:struct	

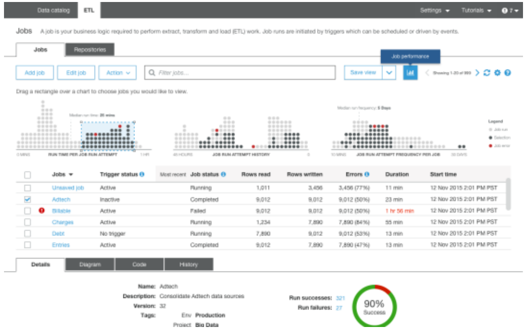
Change	Column	Data type	Default value
	contributors	null	
	created_at	string	
Added	delete	▸ struct	
	entities	▾ struct	
	hashtags	▸ array:struct	
	media	▸ array:struct	



Script: Version 0 (current) ▾

```
from pysparkcontext import SparkContext
from glue.context import GlueContext
from glue.transforms import Relationalize
from glue.transforms import SelectFromCollection
from glue.transforms import DropNullFields

sc = SparkContext()
glueContext = GlueContext(sc)
## @type: DataSource
## @args: [name_space = "glue-demo-25", table_name =
    "simple_tweets_json_3c62f827"]
## @return: datasource0
## @inputs: []
datasource0 = glueContext.create_dynamic_frame.from_catalog(name_space = "glue
    -demo-25", table_name = "simple_tweets_json_3c62f827")
## @type: Relationalize
## @args: [name = "simple_tweets_json_3c62f827"]
## @return: relationalize1
## @inputs: [frame = datasource0]
relationalize1 = Relationalize.apply(frame = datasource0, name =
    "simple_tweets_json_3c62f827")
## @type: SelectFromCollection
## @args: [key = "simple_tweets_json_3c62f827_entities.hashtags"]
## @return: selectfromcollection4
## @inputs: [dfc = relationalize1]
selectfromcollection4 = SelectFromCollection.apply(dfc = relationalize1, key =
    "simple_tweets_json_3c62f827_entities.hashtags")
## @type: DropNullFields
## @args: []
## @return: dropnullfields5
## @inputs: [frame = selectfromcollection4]
dropnullfields5 = DropNullFields.apply(frame = selectfromcollection4)
## @type: DataSink
## @args: [catalog_connection = "Redshift2", connection_options = {"dbtable":
    "simple_tweets_json_3c62f827_entities.hashtags", "database": "testdb"}]
## @return: datasink6
## @inputs: [frame = dropnullfields5]
datasink6 = glueContext.write_dynamic_frame.from_jdbc_conf(frame =
    dropnullfields5, catalog_connection = "Redshift2", connection_options =
```



Jobs A job is your business logic required to perform extract, transform and load (ETL) work. Job runs are initiated by triggers which can be scheduled or driven by events.

Drag a rectangle over a chart to choose jobs you would like to view.

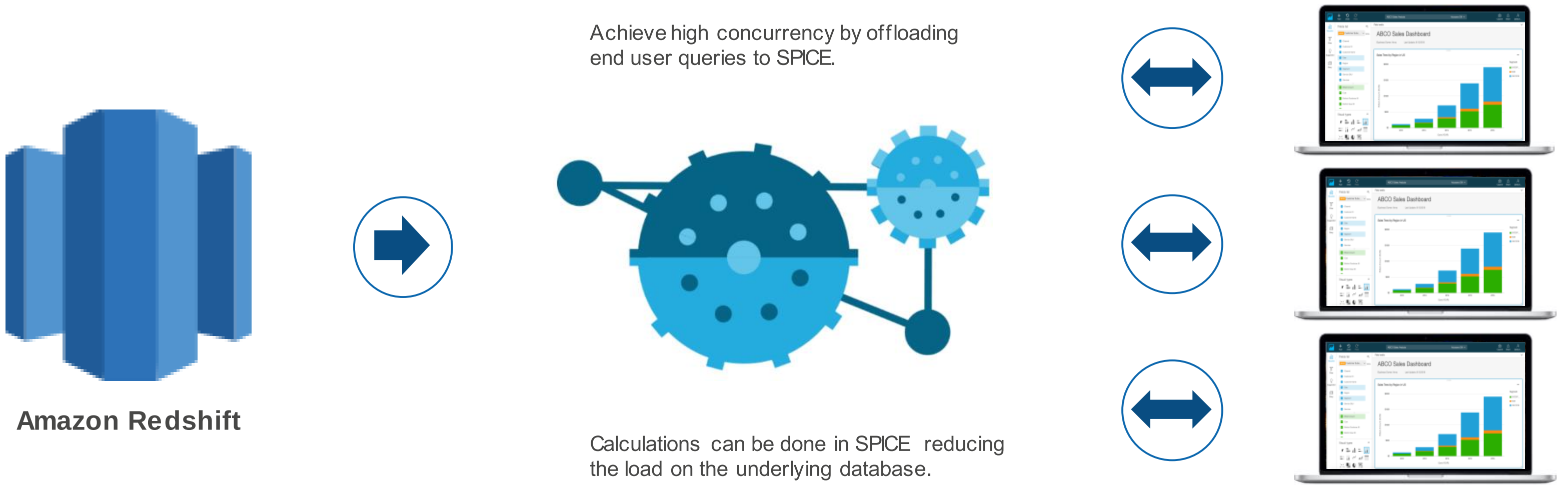
□		Jobs ▾	Trigger status	Most recent Job status	Rows read	Rows written	Errors	Duration	Start time
□		Unsaved job	Active	Running	1,011	3,456	3,456 (77%)	11 min	12 Nov 2015 2:01 PM PST
☑		Adtech	Inactive	Completed	9,012	9,012	9,012 (50%)	23 min	12 Nov 2015 2:01 PM PST
□	!	Billable	Active	Failed	9,012	9,012	9,012 (50%)	1 hr 56 min	12 Nov 2015 2:01 PM PST
□		Charges	Active	Running	1,234	7,890	7,890 (84%)	55 min	12 Nov 2015 2:01 PM PST
□		Debt	No trigger	Running	7,890	9,012	9,012 (53%)	13 min	12 Nov 2015 2:01 PM PST
□		Entries	Active	Completed	9,012	7,890	7,890 (47%)	13 min	12 Nov 2015 2:01 PM PST

Name: Adtech
Description: Consolidate Adtech data sources
Version: 32
Tags: Env **Production**
Project **Big Data**

Run successes: 321
Run failures: 27

QuickSight for BI on Redshift

QuickSight seamlessly connects to Redshift giving you native access to all of your clusters and tables.



Achieve high concurrency by offloading end user queries to SPICE.

Calculations can be done in SPICE reducing the load on the underlying database.

Amazon Redshift Spectrum

Amazon Redshift Spectrum

Run SQL queries directly against data in S3 using thousands of nodes

Fast @ exabyte scale

Elastic & highly available

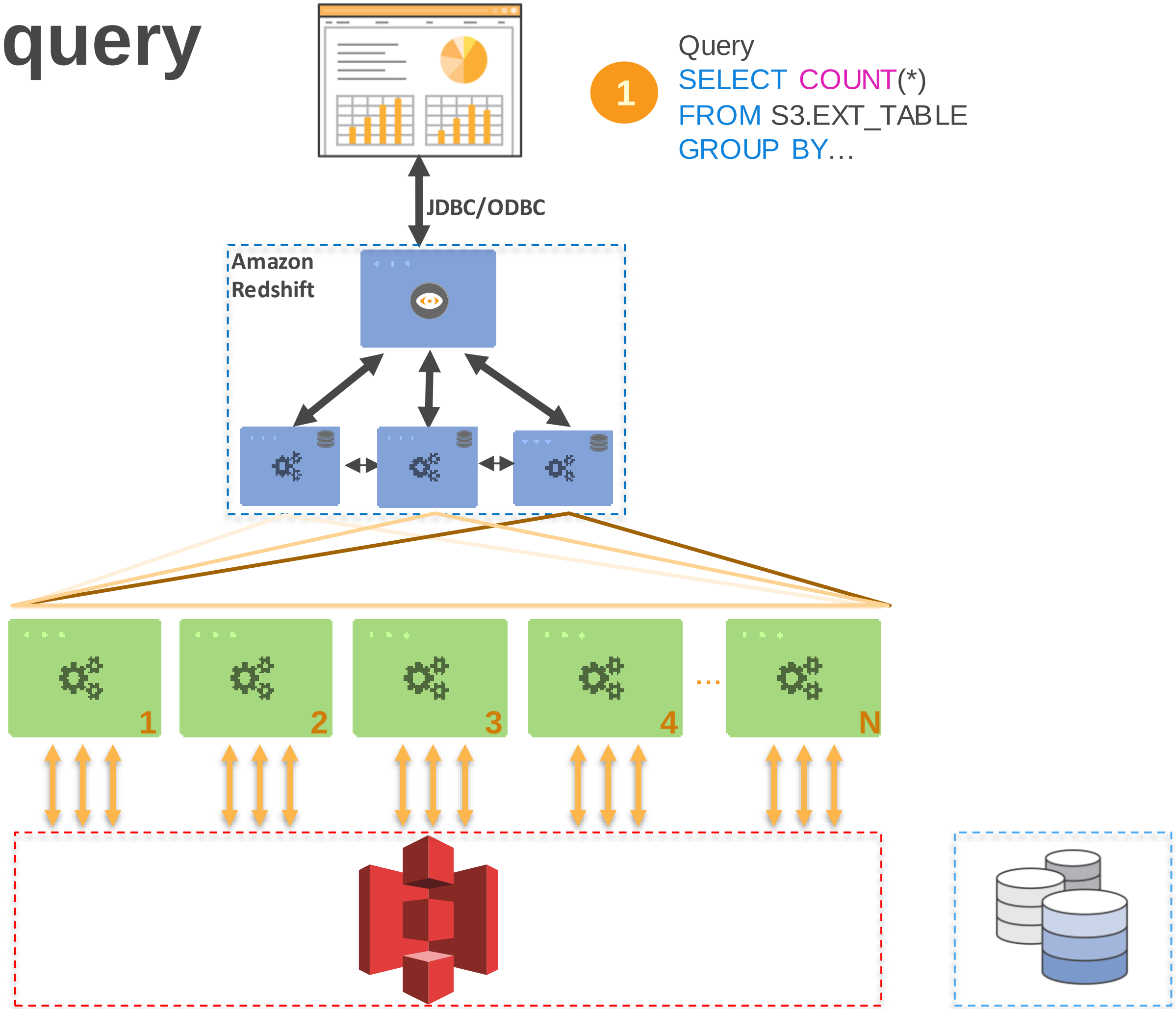
On-demand, pay-per-query

High concurrency: Multiple clusters access same data

No ETL: Query data in-place using open file formats

Full Amazon Redshift SQL support

Life of a query



Query

```
SELECT COUNT(*)
FROM S3.EXT_TABLE
GROUP BY…
```

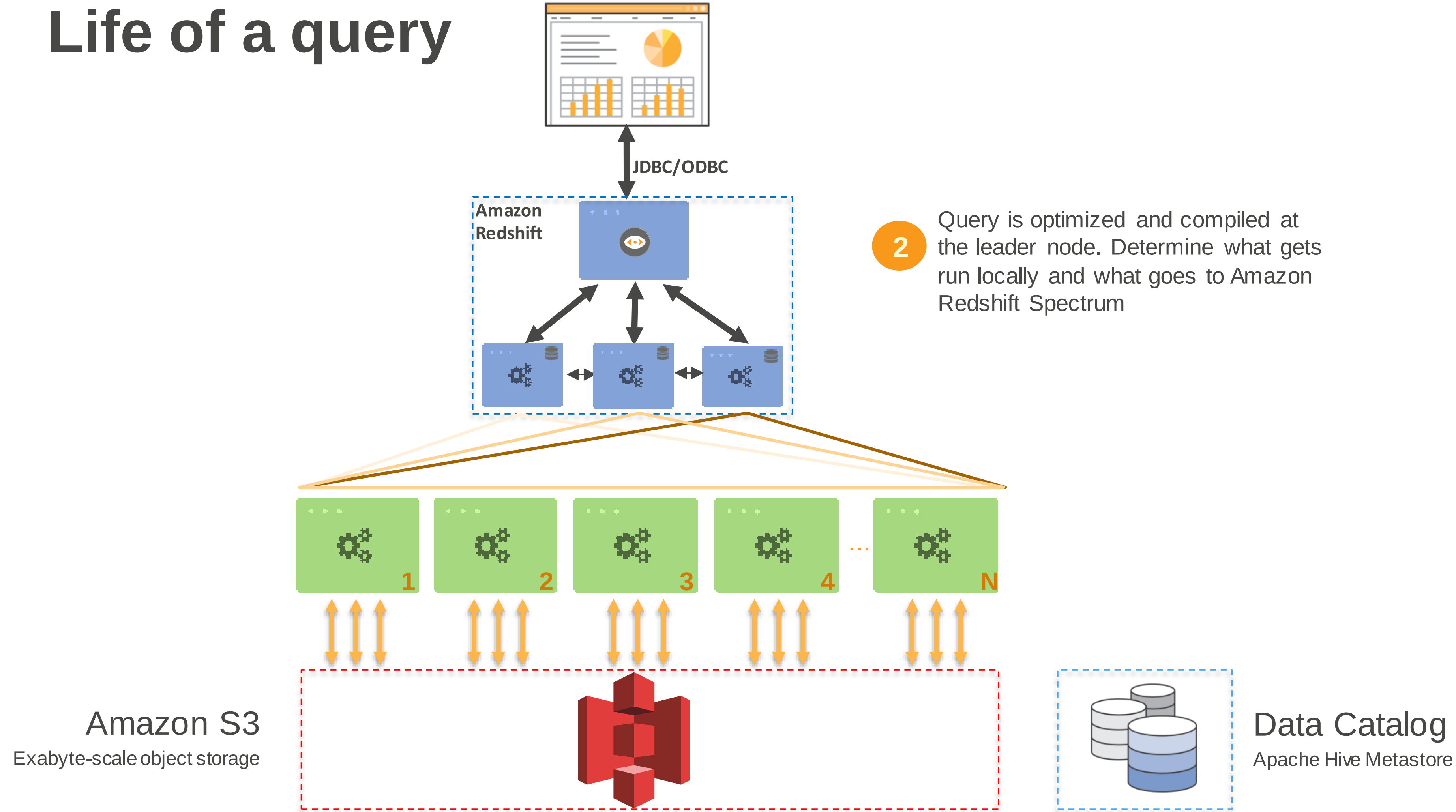
Amazon S3

Exabyte-scale object storage

Data Catalog

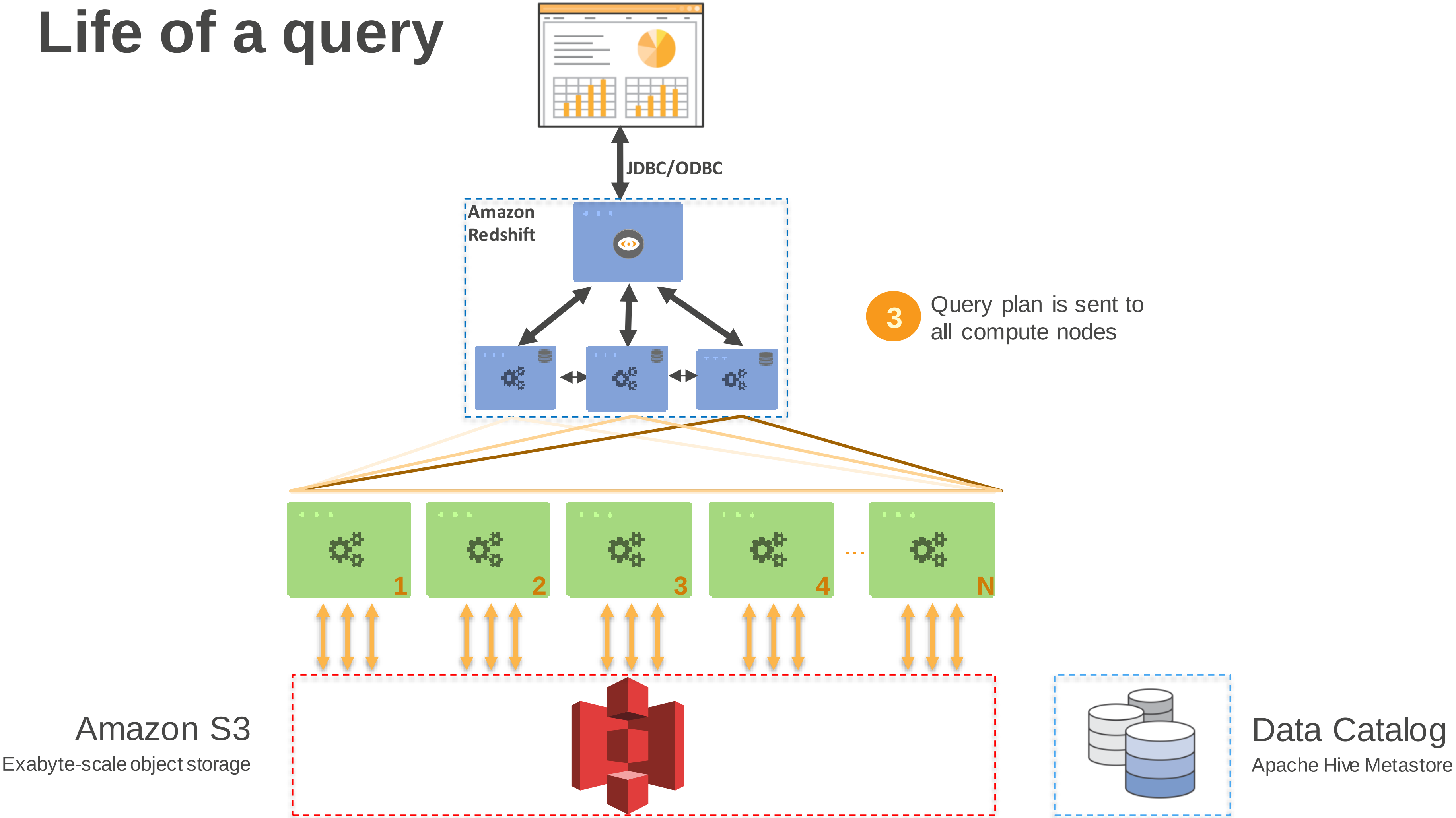
Apache Hive Metastore

Life of a query



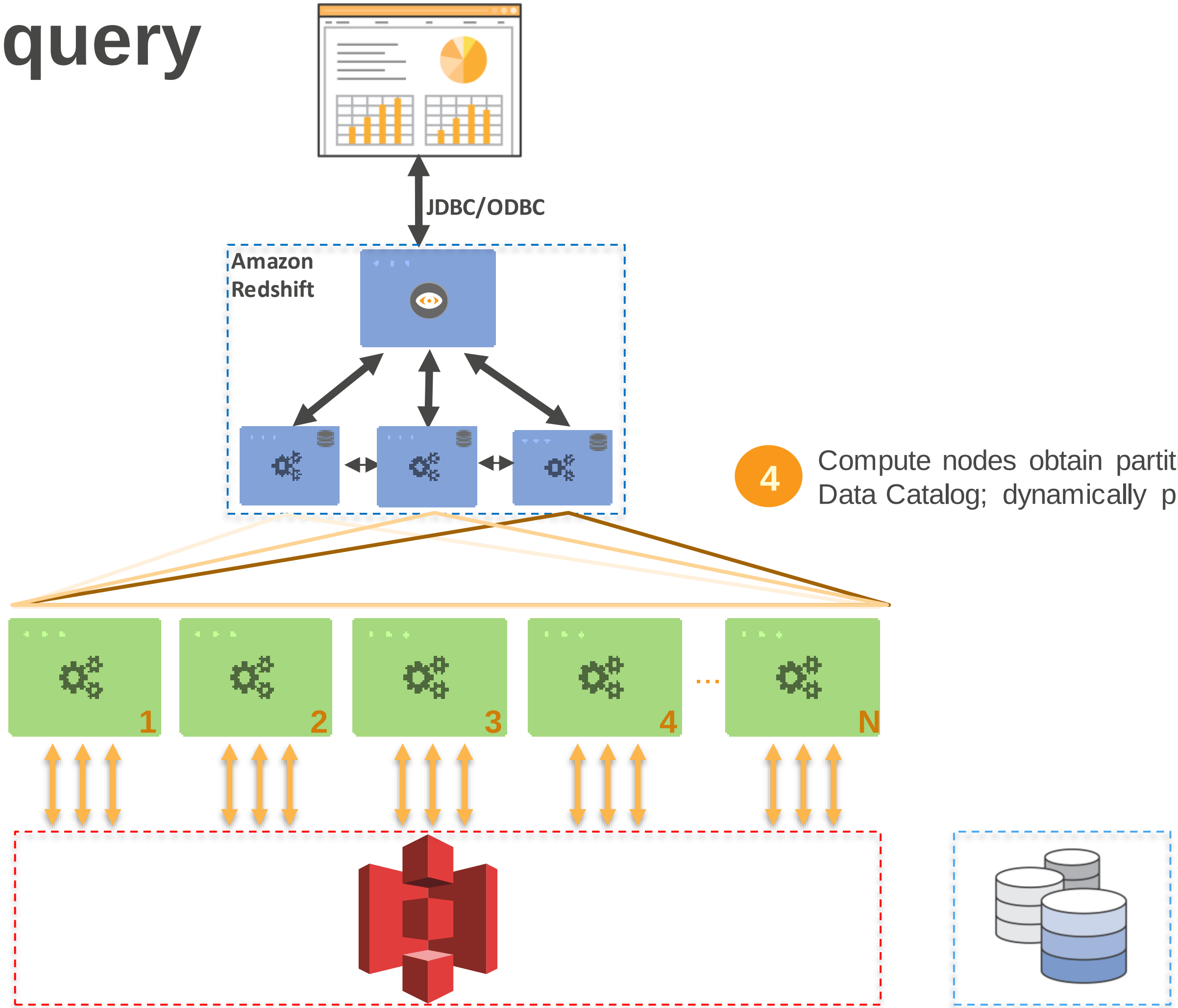
2 Query is optimized and compiled at the leader node. Determine what gets run locally and what goes to Amazon Redshift Spectrum

Life of a query



3 Query plan is sent to all compute nodes

Life of a query

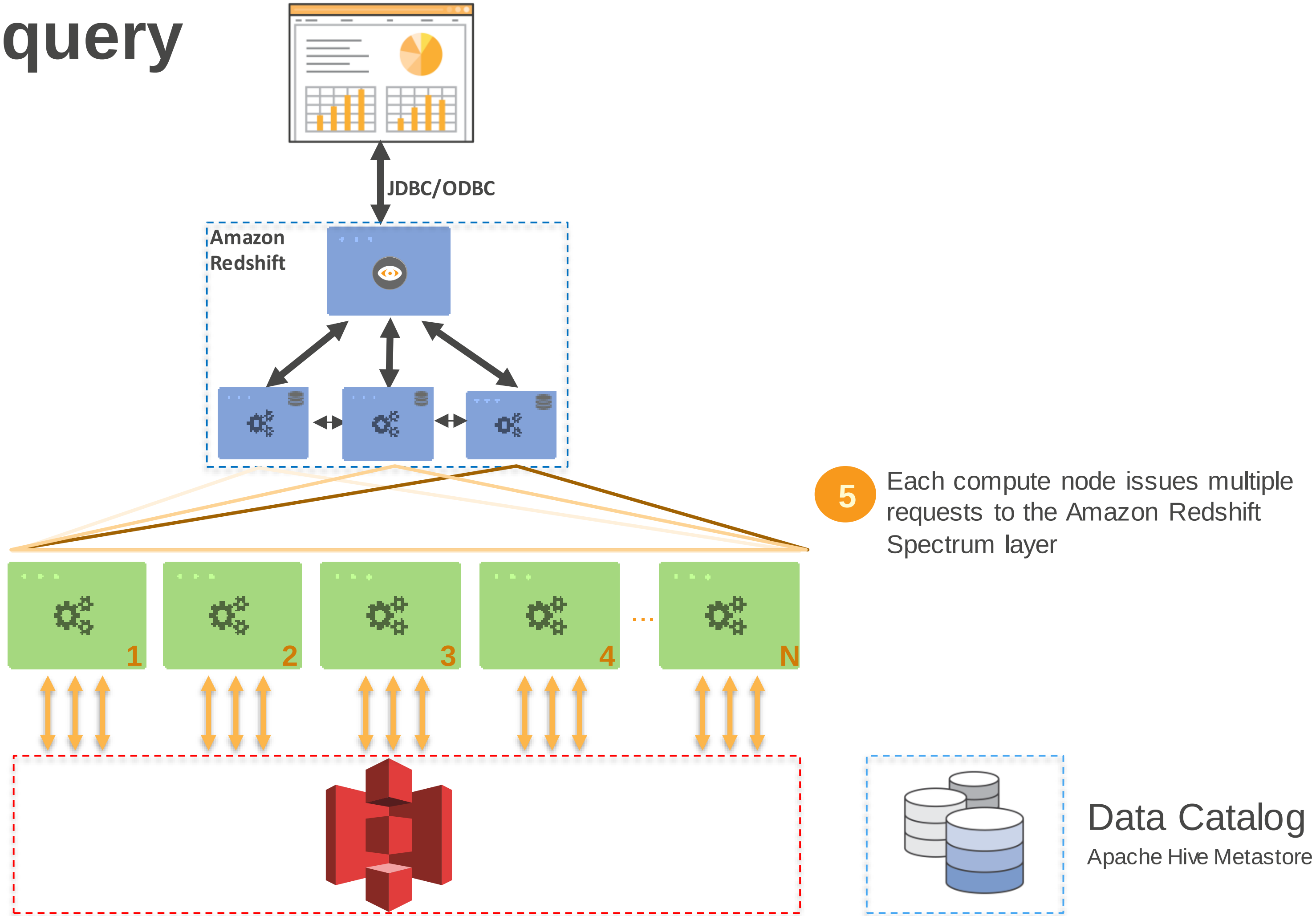


4 Compute nodes obtain partition info from Data Catalog; dynamically prune partitions

Amazon S3
Exabyte-scale object storage

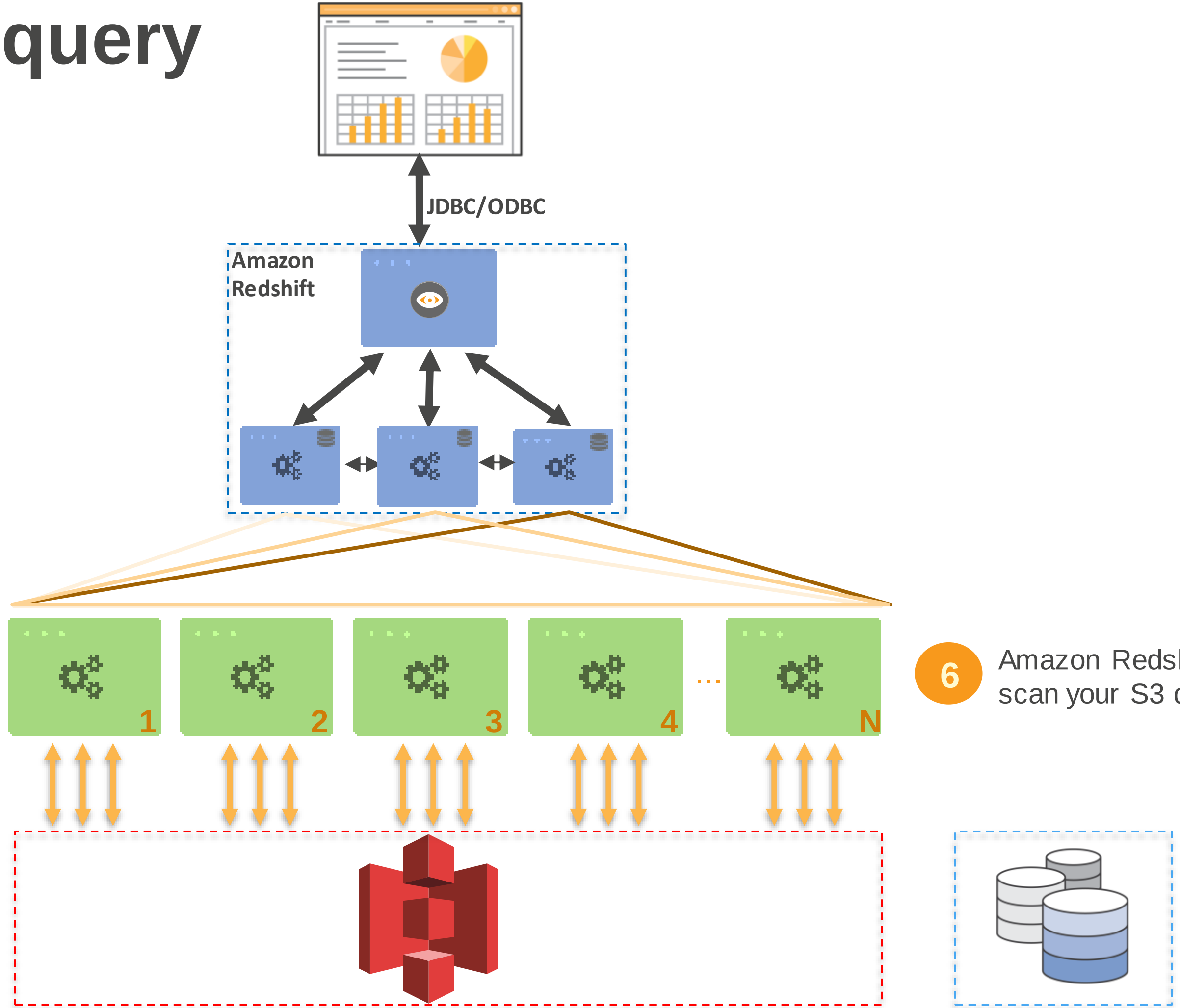
Data Catalog
Apache Hive Metastore

Life of a query



Each compute node issues multiple requests to the Amazon Redshift Spectrum layer

Life of a query



Amazon Redshift Spectrum nodes scan your S3 data

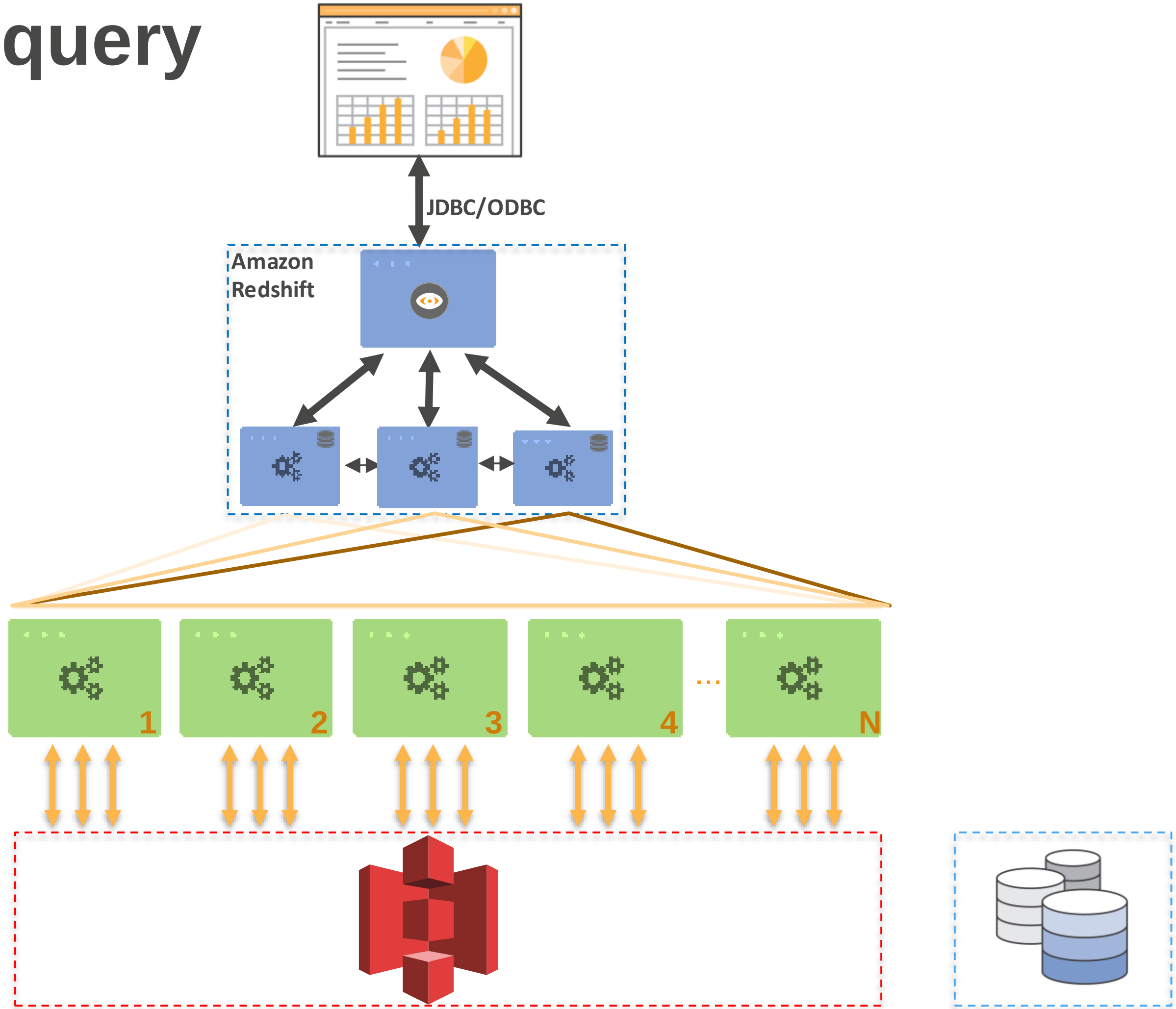
Amazon S3

Exabyte-scale object storage

Data Catalog

Apache Hive Metastore

Life of a query



7 Amazon Redshift Spectrum projects, filters, joins and aggregates

Amazon S3

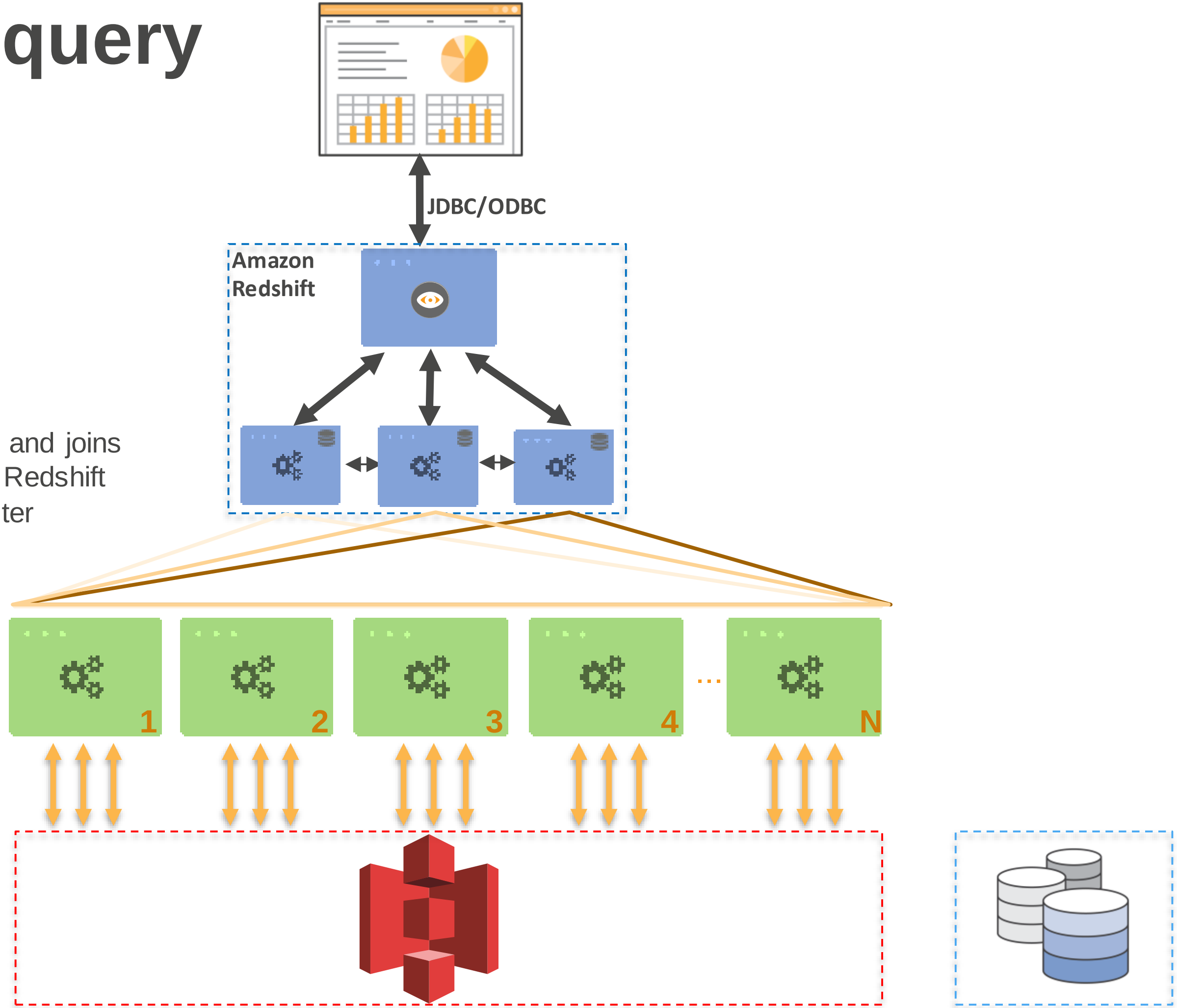
Exabyte-scale object storage

Data Catalog

Apache Hive Metastore

Life of a query

8 Final aggregations and joins with local Amazon Redshift tables done in-cluster



Amazon S3

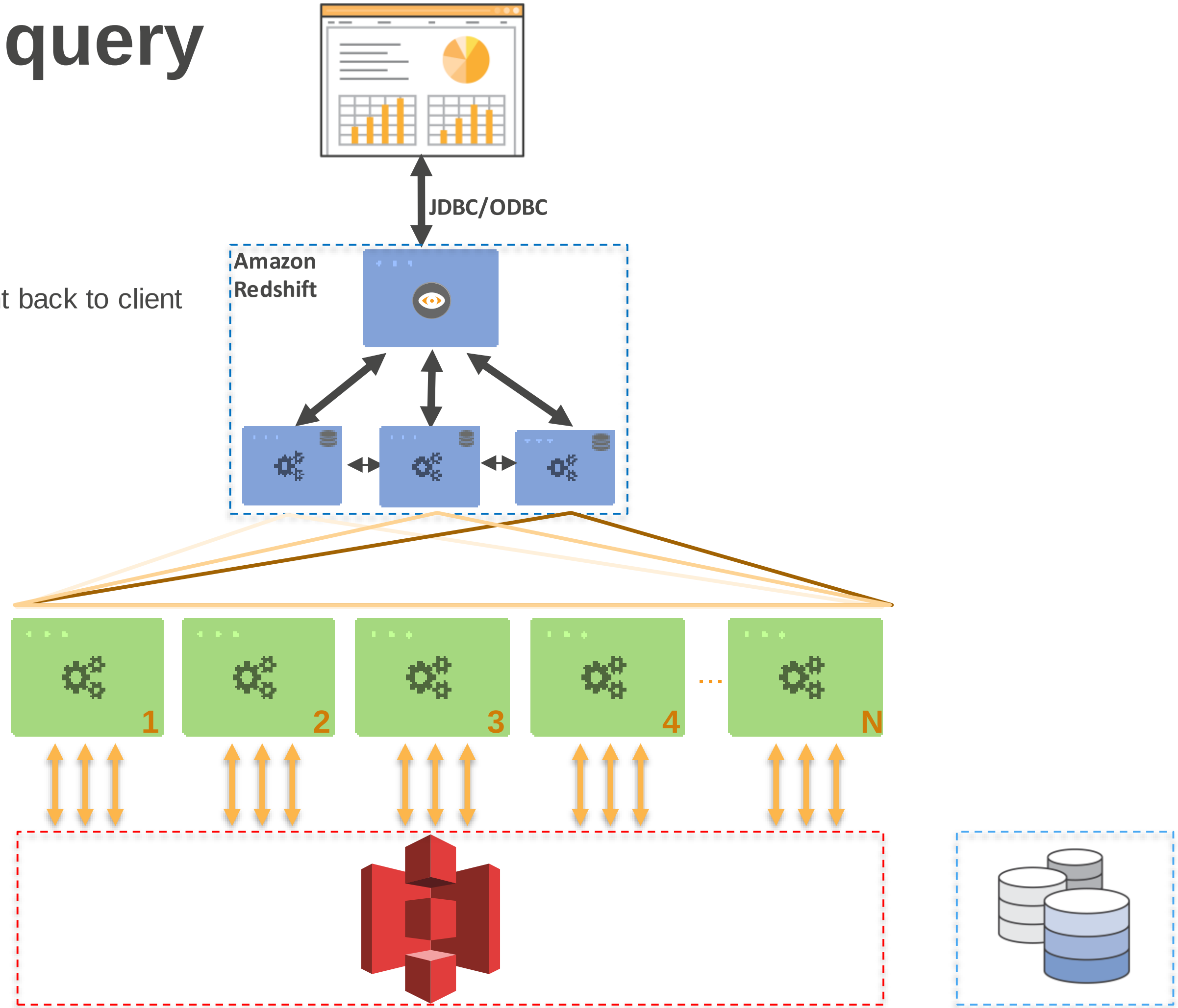
Exabyte-scale object storage

Data Catalog

Apache Hive Metastore

Life of a query

9 Result is sent back to client



Amazon S3

Exabyte-scale object storage

Data Catalog

Apache Hive Metastore

Demo:
Running an analytic query over an exabyte in S3

Lets build an analytic query - #1

An author is releasing the 8th book in her popular series. How many should we order for Seattle? What were prior first few day sales?

Lets get the prior books she’s written.

1 Table

2 Filters

```
SELECT
       P.ASIN,
       P.TITLE
FROM
       products P
WHERE
       P.TITLE LIKE ‘%POTTER%’ AND
       P.AUTHOR = ‘J. K. Rowling’
```

Lets build an analytic query - #2

An author is releasing the 8th book in her popular series. How many should we order for Seattle? What were prior first few day sales?

Lets compute the sales of the prior books she’s written in this series and return the top 20 values

2 Tables (1 S3, 1 local)

2 Filters

1 Join

2 Group By columns

1 Order By

1 Limit

1 Aggregation

```
SELECT
       P.ASIN,
       P.TITLE,
       SUM(D.QUANTITY * D.OUR_PRICE) AS SALES_sum
FROM
       s3.d_customer_order_item_details D,
       products P
WHERE
       D.ASIN = P.ASIN AND
       P.TITLE LIKE '%Potter%' AND
       P.AUTHOR = 'J. K. Rowling' AND
GROUP BY P.ASIN, P.TITLE
ORDER BY SALES_sum DESC
LIMIT 20;
```

Lets build an analytic query - #3

An author is releasing the 8th book in her popular series. How many should we order for Seattle? What were prior first few day sales?

Lets compute the sales of the prior books she's written in this series and return the top 20 values, just for the first three days of sales of first editions

3 Tables (1 S3, 2 local)

5 Filters

2 Joins

3 Group By columns

1 Order By

1 Limit

1 Aggregation

1 Function

2 Casts

```
SELECT
       P.ASIN,
       P.TITLE,
       P.RELEASE_DATE,
       SUM(D.QUANTITY * D.OUR_PRICE) AS SALES_sum
FROM
       s3.d_customer_order_item_details D,
       asin_attributes A,
       products P
WHERE
       D.ASIN = P.ASIN AND
       P.ASIN = A.ASIN AND
       A.EDITION LIKE '%FIRST%' AND
       P.TITLE LIKE '%Potter%' AND
       P.AUTHOR = 'J. K. Rowling' AND
       D.ORDER_DAY :: DATE >= P.RELEASE_DATE AND
       D.ORDER_DAY :: DATE < dateadd(day, 3, P.RELEASE_DATE)
GROUP BY P.ASIN, P.TITLE, P.RELEASE_DATE
ORDER BY SALES_sum DESC
LIMIT 20;
```

Lets build an analytic query - #4

An author is releasing the 8th book in her popular series. How many should we order for Seattle? What were prior first few day sales?

Lets compute the sales of the prior books she’s written in this series and return the top 20 values, just for the first three days of sales of first editions in the city of Seattle, WA, USA

4 Tables (1 S3, 3 local)

8 Filters

3 Joins

4 Group By columns

1 Order By

1 Limit

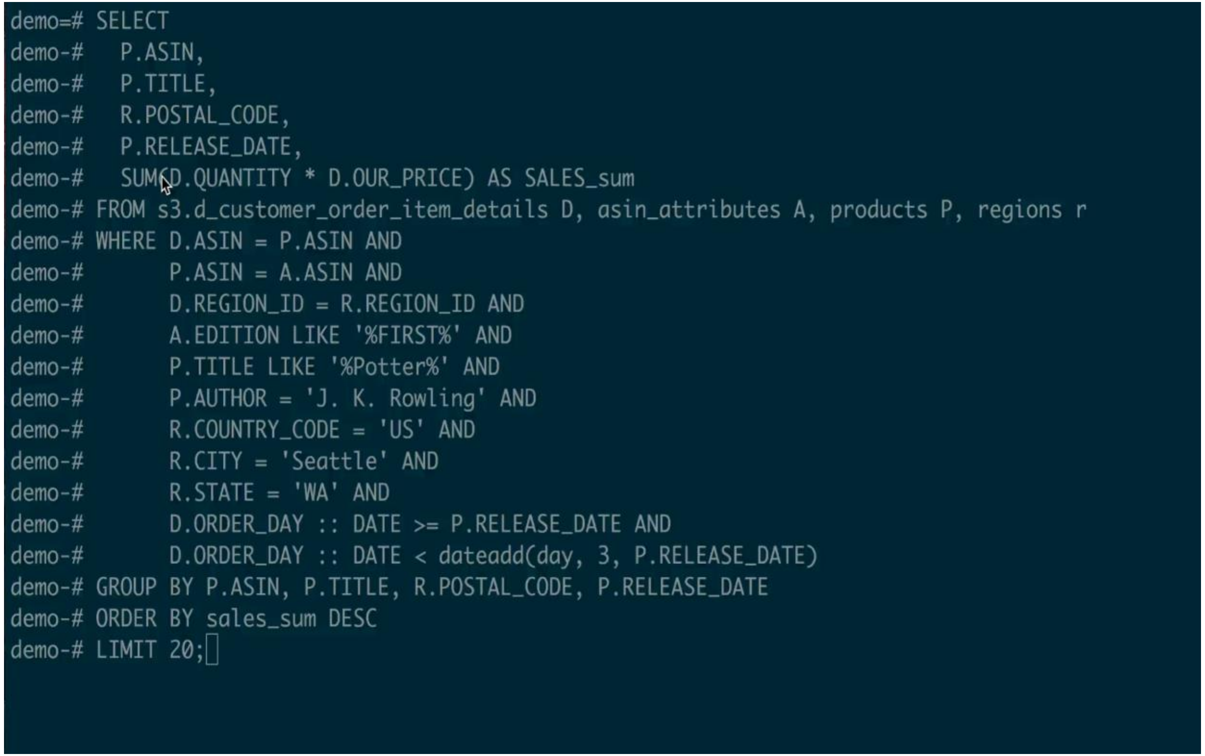
1 Aggregation

1 Function

2 Casts

```
SELECT
       P.ASIN,
       P.TITLE,
       R.POSTAL_CODE,
       P.RELEASE_DATE,
       SUM(D.QUANTITY * D.OUR_PRICE) AS SALES_sum
FROM
       s3.d_customer_order_item_details D,
       asin_attributes A,
       products P,
       regions R
WHERE
       D.ASIN = P.ASIN AND
       P.ASIN = A.ASIN AND
       D.REGION_ID = R.REGION_ID AND
       A.EDITION LIKE '%FIRST%' AND
       P.TITLE LIKE '%Potter%' AND
       P.AUTHOR = 'J. K. Rowling' AND
       R.COUNTRY_CODE = ‘US’ AND
       R.CITY = ‘Seattle’ AND
       R.STATE = ‘WA’ AND
       D.ORDER_DAY :: DATE >= P.RELEASE_DATE AND
       D.ORDER_DAY :: DATE < dateadd(day, 3, P.RELEASE_DATE)
GROUP BY P.ASIN, P.TITLE, R.POSTAL_CODE, P.RELEASE_DATE
ORDER BY SALES_sum DESC
LIMIT 20;
```

Now let’s run that query over an exabyte of data in S3



```
demo=# SELECT
demo-#   P.ASIN,
demo-#   P.TITLE,
demo-#   R.POSTAL_CODE,
demo-#   P.RELEASE_DATE,
demo-#   SUM(D.QUANTITY * D.OUR_PRICE) AS SALES_sum
demo-# FROM s3.d_customer_order_item_details D, asin_attributes A, products P, regions r
demo-# WHERE D.ASIN = P.ASIN AND
demo-#       P.ASIN = A.ASIN AND
demo-#       D.REGION_ID = R.REGION_ID AND
demo-#       A.EDITION LIKE '%FIRST%' AND
demo-#       P.TITLE LIKE '%Potter%' AND
demo-#       P.AUTHOR = 'J. K. Rowling' AND
demo-#       R.COUNTRY_CODE = 'US' AND
demo-#       R.CITY = 'Seattle' AND
demo-#       R.STATE = 'WA' AND
demo-#       D.ORDER_DAY :: DATE >= P.RELEASE_DATE AND
demo-#       D.ORDER_DAY :: DATE < dateadd(day, 3, P.RELEASE_DATE)
demo-# GROUP BY P.ASIN, P.TITLE, R.POSTAL_CODE, P.RELEASE_DATE
demo-# ORDER BY sales_sum DESC
demo-# LIMIT 20;
```

Roughly 140 TB of customer item order detail records for each day over past 20 years.

190 million files across 15,000 partitions in S3. One partition per day for USA and rest of world.

Need a **billion-fold** reduction in data processed.

Running this query using a 1000 node Hive cluster would take over 5 years.*

- Compression ……………..……..……..5X
- Columnar file format……………..…10X

- Scanning with 2500 nodes……..2500X

- Static partition elimination…………….2X
- Dynamic partition elimination..…..350X
- Redshift’s query optimizer………….40X

Total reduction………..………..…**3.5B X**

* Estimated using 20 node Hive cluster & 1.4TB, assume linear
* Query used a 20 node DC1.8XLarge Amazon Redshift cluster
* Not actual sales data - generated for this demo based on data format used by Amazon Retail.

Is Amazon Redshift Spectrum useful if I don’t have an exabyte?

Your data will get bigger

- On average, data warehousing volumes grow 10x every 5 years
- The average Amazon Redshift customer doubles data each year

Amazon Redshift Spectrum makes data analysis simpler

- Access your data without ETL pipelines
- Teams using Amazon EMR, Athena & Redshift can collaborate using the same data lake

Amazon Redshift Spectrum improves availability and concurrency

- Run multiple Amazon Redshift clusters against common data
- Isolate jobs with tight SLAs from ad hoc analysis

Redshift Partner Echo System

4 types of partners

- Load and transform your data with Data Integration Partners
- Analyze data and share insights across your organization with Business Intelligence Partners
- Architect and implement your analytics platform with System Integration and Consulting Partners
- Query, explore and model your data using tools and utilities from Query and Data Modeling Partners

aws.amazon.com/redshift/partners/

“Some” Amazon Redshift Customers

Adobe
BEACHMINT
yelp
SNOWPLOW
Nintendo
AMPLITUDE
NOKIA
foursquare
Pinterest
FT.com FINANCIAL TIMES
sling
latentview Actionable Insights • Accurate Decisions
NTT docomo
NASDAQ OMX
FINRA
amazon
etix
4
scopely
has offers
imshealth INTELLIGENCE APPLIED.
euclid
SOUNDCLOUD
Sansan
Schumachergroup
Albert Optimization technology
spuul
peak GAMES
CAKE
BookmyShow
VIVAKI
DataXu
MINICLIP
UMUC University of Maryland University College

Manchester Airport Group

An AWS Redshift customer story

MAG – take-off with cloud and data

Stuart Hutson – Head of Data and BI

THE AVIATION PROFESSIONALS

MAG is a leading UK based airport company, which owns and operates Manchester, London Stansted, East Midlands and Bournemouth airports.

MAG is privately managed on behalf of its shareholders, the local authorities of Greater Manchester and Industry Funds Management (IFM). IFM is a highly experienced, long-term investor in airports and already has significant interests in ten airports across Australia and Europe.

MANCHESTER CITY COUNCIL

AGMA ASSOCIATION OF GREATER MANCHESTER AUTHORITIES

Voting: 50%
Economic: 64.5%

ifm investors

Voting: 50%
Economic: 35.5%

48.5 MILLION passengers served per year.

Over **80 AIRLINES** serving **272 DESTINATIONS** direct.

£134.3 MILLION RETAIL INCOME per annum delivered via 200+ shops, bars and restaurants.

£125.7 MILLION CAR PARKS INCOME delivered via 96,000 parking spaces.

£623 MILLION property assets across all airports, 5.67m sq ft of commercial property.

£738.4 MILLION REVENUE +10.0% increase from last year.

£283.6 MILLION EBITDA growth of 17.2% in 2015.

£5.6 BILLION contribution to the UK economy from MAG airports.

manchester airport PART OF M.A.G

MAG USA

london stansted airport PART OF M.A.G

east midlands airport PART OF M.A.G

bournemouth airport PART OF M.A.G

Source: FY15 Report & Accounts

MAG

OUR AIRPORTS…

MAG airports serve over 48.5 million people per annum from complementary catchment areas covering over 75% of the UK population.

manchester airport

PART OF M.A.G

c. 23m passengers per annum.

UK's 3rd largest airport.

70+ airlines & 200+destinations.

2 runways with potential 62% capacity.

21.5m people within a 2 hour drive.

bournemouth airport

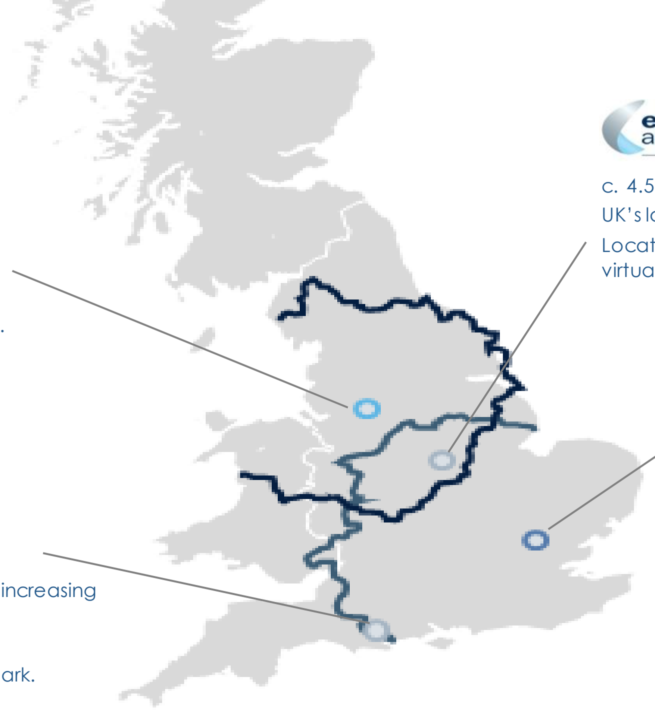
PART OF M.A.G

c. 0.7m passengers per annum.

Significant investment in new terminal increasing passenger capacity to 3m p.a.

Wealthy catchment area.

Large land holding – on-site business park.



east midlands airport

PART OF M.A.G

c. 4.5m passengers per annum.

UK's largest freight airport after Heathrow – 310,000 tonnes p.a.

Located next to key road interchanges – four hours from virtually all UK commerce.

london stansted airport

PART OF M.A.G

c. 23m passengers per annum.

UK's 4th largest airport.

150+ destinations.

1 runway with 50% spare capacity.

25m people within 2 hour drive.

Acquired February 2015.

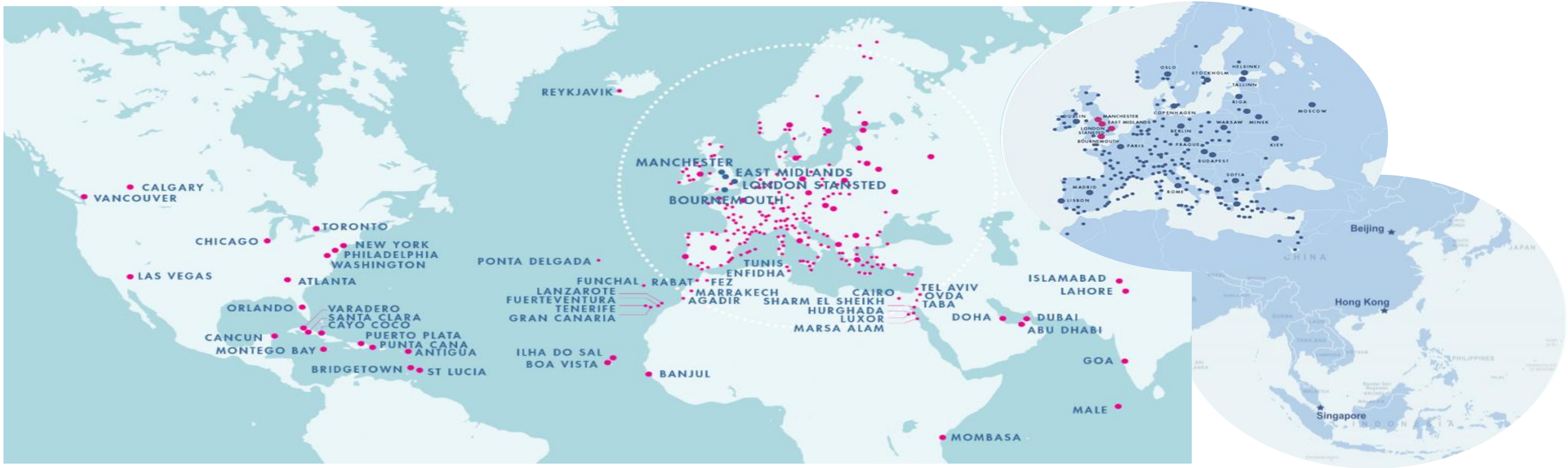
Catchment area within 2 hours' drive of:

MAN

STN

OUR CONNECTIVITY…

80+ airlines and over 270 direct destinations providing global connectivity.



AIR SERVICE DEVELOPMENT

MAG has a diverse carrier mix from global destinations with an excellent track record of incentivizsng passenger growth.

MAG has exceeded expectations with industry leading rates of passenger growth. Importantly for passengers, by forging strong commercial partnerships with airlines, our airports have been able to increase choice and convenience and make a stronger contribution to economy growth.

CARGO SERVICE DEVELOPMENT

MAG's Cargo produces an annual income of £20.2 million and holds 26% of the UK freight market share.

East Midlands is the UK's largest dedicated freight hub handling 310,000 tonnes of freight per annum. Stansted handles 233,000 tonnes of freight per annum and is a key gateway to London and the South of England.

BRITISH AIRWAYS CATHAY PACIFIC RYANAIR Emirates American Airlines 海南航空 HAINAN AIRLINES DELTA KLM الاتحاد ETIHAD AIRWAYS MAG

OUR DEVELOPMENTS…

Manchester Transformation Programme and London Stansted Transformation Programme are developments that all aim to drive improved customer service.

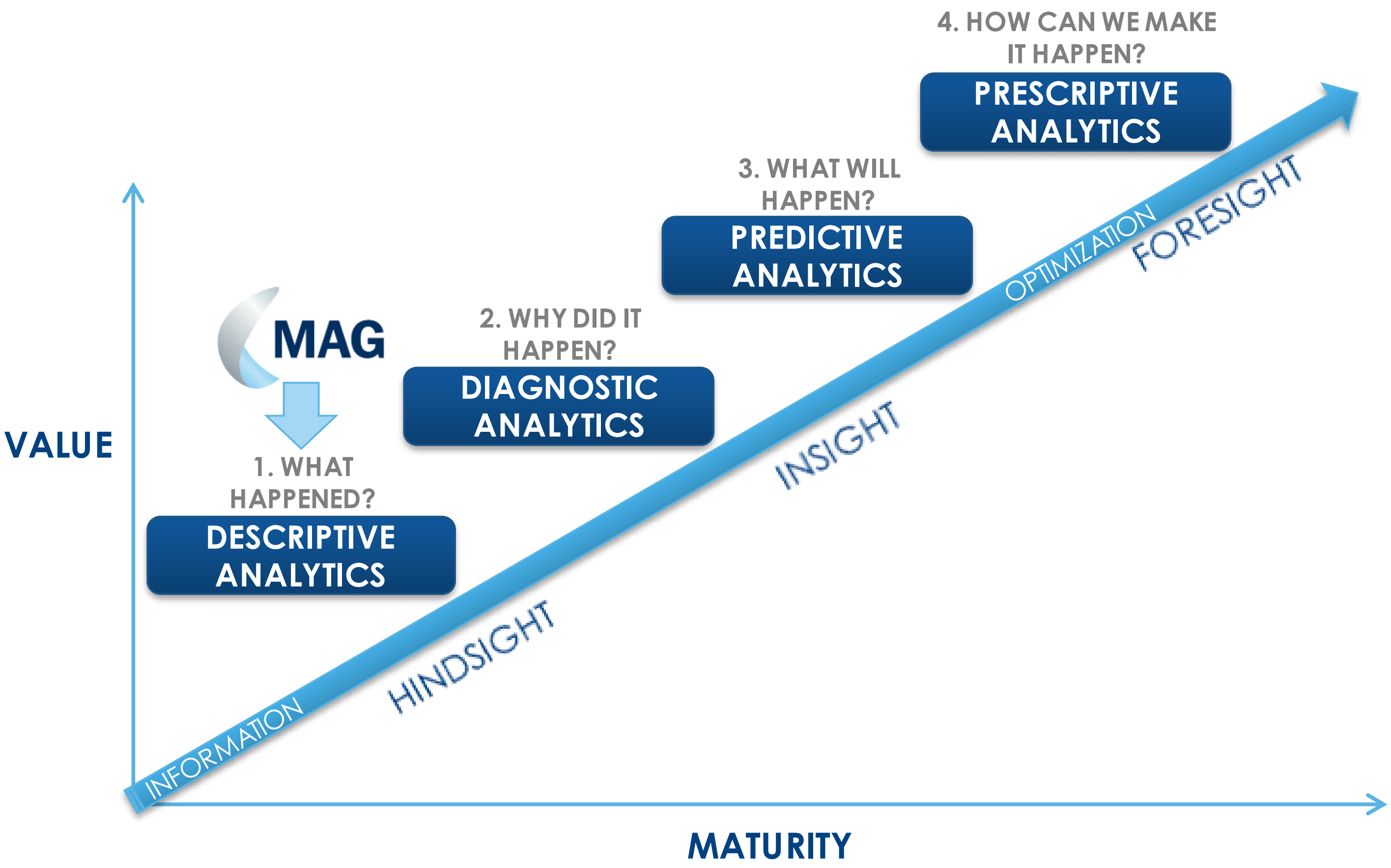
MANCHESTER TRANSFORMATION PROGRAMME

With investment of £1 billion, Manchester will become one of the most modern and customer focused airports in Europe demonstrating the importance of Manchester as a global gateway.

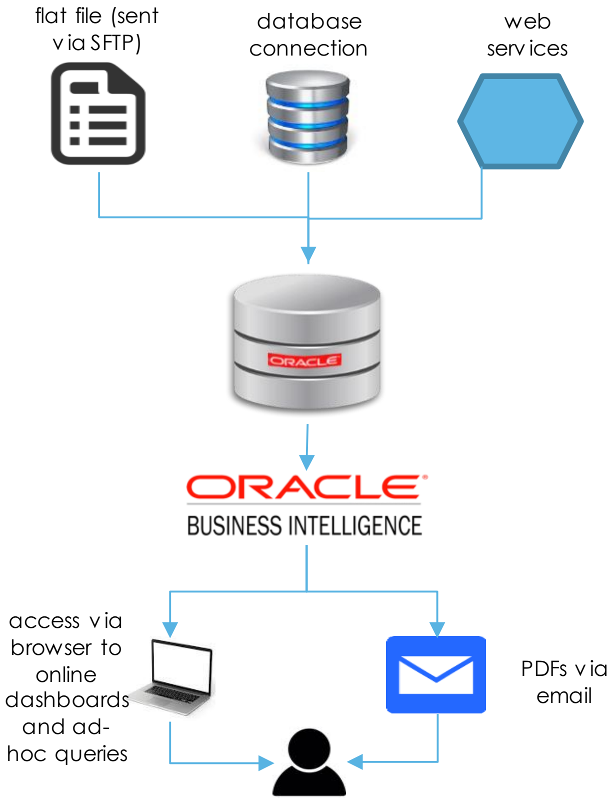
LONDON STANSTED TRANSFORMATION PROGRAMME

The £80 million terminal transformation project at London Stansted will transform the passenger experience and boost commercial yields.

MAG'S CURRENT BUSINESS INTELLIGENCE MATURITY



MAG’S LEGACY ARCHITECTURE - CHALLENGES…


flat file (sent via SFTP)
database connection
web services
ORACLE
ORACLE BUSINESS INTELLIGENCE
access via browser to online dashboards and ad-hoc queries
PDFs via email

Business

- Multiple version of the truth
- Unclear of operational problems
- People are overloaded and with data and data led questions
- Analysts not able to do analyst job due to lack of data and tools
- Data processing issues - late reports, missing data
- Data accessible in silos - no real cross-functional analysis

Technical

- Database @ 95% capacity on physical kit that can not be scaled.
- Dashboards are slow to run.
- Constant optimisation and maintenance of database.
- Limited concurrent connections for queries.
- Lack of self-serve – centralised BI model.
- No direct connection to database – business wants to expand into using R and Python etc.
- All data in batch with no possibility of streaming

VALUE OF BI STRATEGY IN MAG…

Monetise Data

- ***Monetise data and technology across our omni-channels:*** *MAG's BI Strategy must be bold, it should be aiming for how we monetises our data and technology across our omni-channel business by improving the customer experience.*

Democratise Data

- ***Democratise data across the Enterprise:*** *Our data needs to be pervasive across the organisation. The decisions of the organisation should be made on clear information presented to the business at the right time to enable MAG to make the right decisions.*

Data DNA

- ***Create a data DNA:*** *Build a culture around data and analytical thinking across the organisation by embedding analytics and data across MAGs business processes and decision making.*

PHASE 1 - IMPLEMENT SELF-SERVE RETAIL BI SOLUTION - 50+ PARTNERS GENERATING OVER £130M REVENUE…

Build Data and BI Foundation solution

- To create an extensible and flexible data solution for MAG comprising of:
    - **Extended Data Warehouse.**
        - Scalable and elastic compute.
        - Deal with seasonality spikes of passenger travel.
    - **Real-time streaming.**
        - Enable MAG to become a real-time business across their customer journey.
    - **Cloud environment:**
        - Secured.
        - Resilient.
        - Repeatable build.
- **Enable MAG to quickly experiment at low cost and minimal risk.**
    - MAG wants to trial new technologies, especially open-source.
- **Create an architecture than can evolve over time to meets MAG's new challenges.**
    - Benefits delivered early and continuously.
    - No need for MAG to invest in a large, front-loaded EDW programme.

EXAMPLE OF MAG'S DESIGN PRINCIPLES TO SOLVE THE PROBLEMS…

- **Evolutionary architecture**
- **Infrastructure as Code**
- **Protecting our data**
- Assume for failure
- Data quality is a priority
- Embrace open source for experimentation
- SaaS -> PaaS -> IaaS
- Serverless computing
- Etc.

MAG – OUR 6 MONTH JOURNEY…

From		To
Single instance database.	→	Scale-able Data Warehouse.
Daily sales rung in at store level.	→	Over 90% of all sales automatically ingested at product level.
Car parking - flat files ingested in batch.	→	Ingest and interrogate streaming data directly: • Car park data is being added via Kinesis
Access to database limited to reporting tool.	→	Authorised users can use visualisation and data science tools (e.g. R and Python) of their choice for self-serve analytics
No database writeback for end-users.	→	Sandboxes in Redshift for end user experimentation.

MAG – NEXT 6-12 MONTHS…

- Moving to near-time streaming into Redshift for:
  - Terminal Operations
  - Security Services
  - Car Park Management

- Streaming semi-structured data into Redshift
  - Trialling IoT data streaming
  - Passenger analysis

- Trial AWS Glue and AWS Redshift Spectrum
  - Automated profile and catalogue of data across the enterprise
  - Continuous integration of data into our data warehouse

Who Are Crimson Macaw

Driving customer success by unlocking the value of data.

Competency focused consultancy

Architecture
and
Data Strategy

DevOps

Data Engineering

Enterprise
Data
Solutions

Data
Science

1. Plan

2. Build

3. Action

www.crimsonmacaw.com

crimsonmacaw

Our partners ...

amazon web services Partner Network

CONSULTING PARTNER

3 AWS Big Data Speciality

3 AWS Certified Solutions Architect Associate

2 AWS Certified Developer Associate

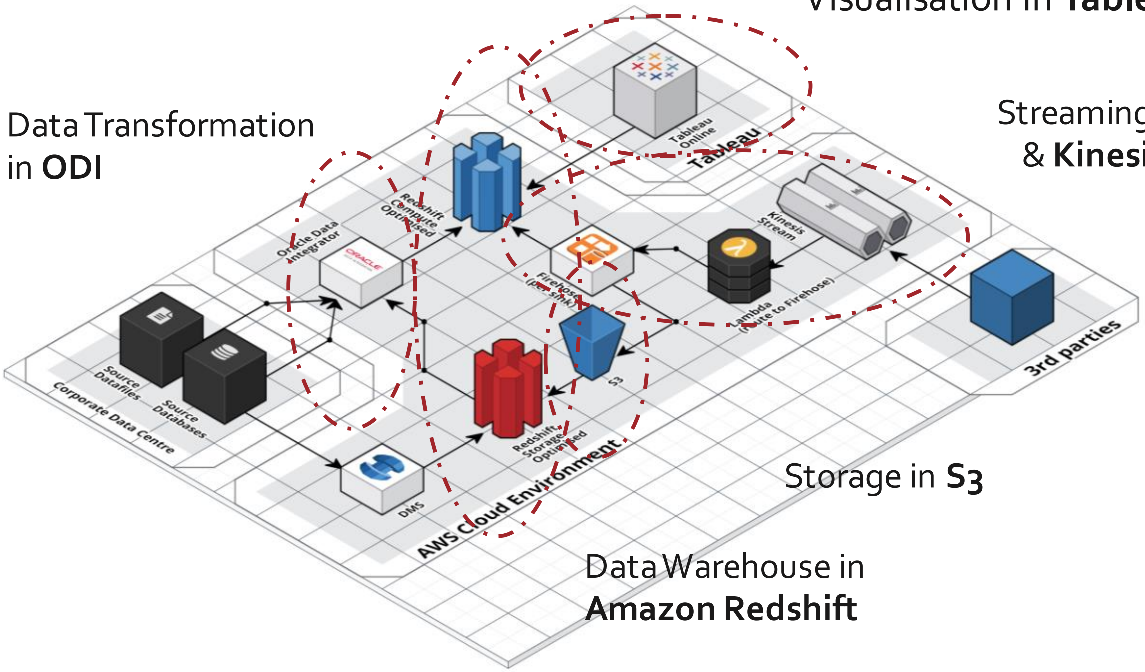
2 AWS Certified SysOps Administrator Associate

tableau

Building a solution

… without too many twists and turns.

Key architectural components used



Visualisation in **Tableau**

Data Transformation in **ODI**

Streaming in **Kinesis** & **Kinesis Firehose**

Storage in **S3**

Data Warehouse in **Amazon Redshift**

crimsonmacaw

Cloud Architecture + Data Architecture = Solution

How do you match the pace of infrastructure build in the cloud with understanding the data & BI requirements?

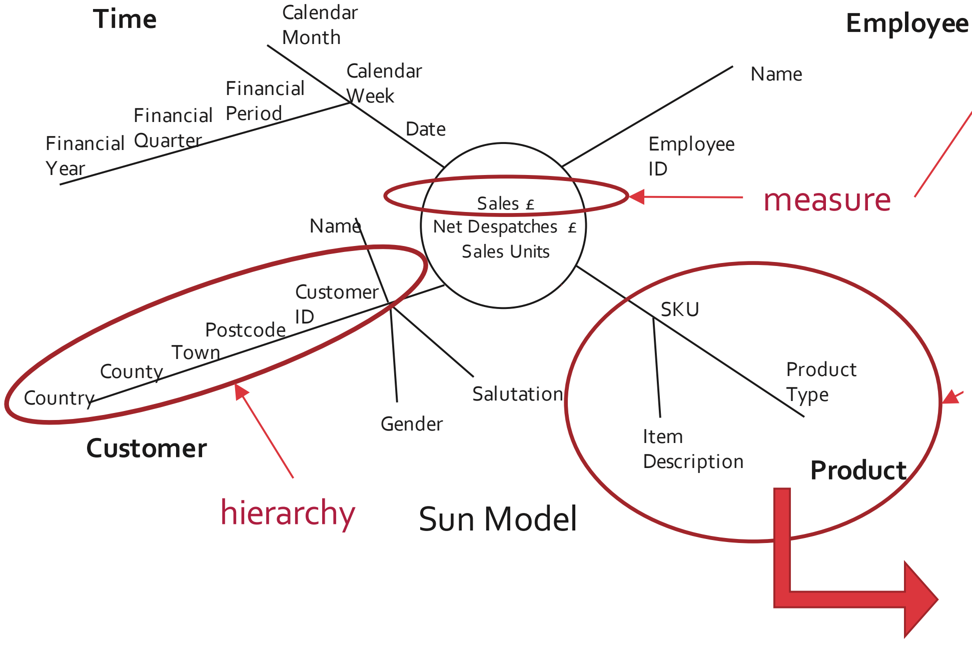
Deliver value quickly vs conformed dimensions?

- A horizontal analytical 'slice' across the estate.
- Understand conformed dimensions.
- Vertical slice of a business domain.
- Reduced refactoring due to the prior horizontal analysis.

Understand how the business will consume and use the data?

- Produce artefacts that are:
  - Shared by stakeholders and the delivery team.
  - Understandable by all parties.
  - Highly visual, allow complex information to be absorbed - sun modelling.

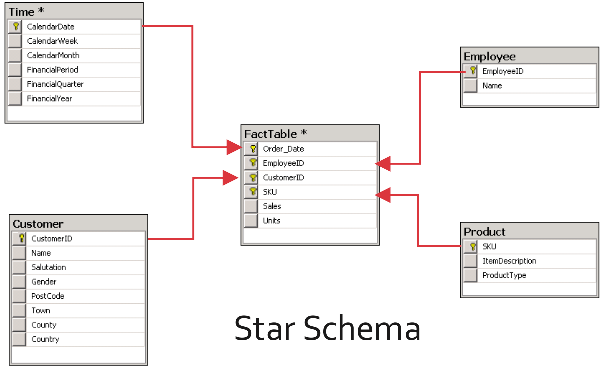
Sun modelling vs Enterprise Bus Matrix



Sun Model

	Dimension					
Measure	Time	Employee	Product	Customer	Return Reason	Subject Area #
Sales £	x	x	x	x		1
Returns £	x	x	x	x	x	2
Net Despatches £	x	x	x	x		1
Sales Units	x	x	x	x		1
Return Units	x	x	x	x	x	2

Enterprise Bus Matrix



Star Schema

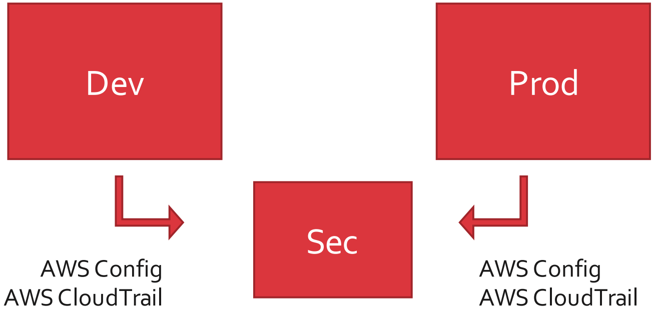
crimsonmacaw

Building the infrastructure (as code)

- Why use infrastructure as code?
  - Repeatability.
  - Consistency.
  - Versioned.
  - Code reviews.
  - Speed of delivery.
- Technology Used:
  - CloudFormation in YAML format with custom YAML Tags.
  - Lambda Functions for Custom Resource Types.
  - Bespoke deployment utility.
  - Puppet Standalone in Cloud Init for EC2.
- Why this approach?
  - Enforced Tagging Policy with propagated tags.
  - Custom YAML Tags act as a precompiler for CloudFormation.
  - Not all resources types were available, e.g. DMS.
  - Redshift IAM Roles and Tags – both now available out of the box!.

Security overview

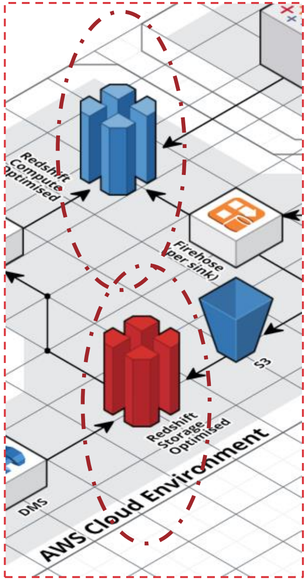
- Three independent AWS accounts
  - Dev – for development of data processes.
  - Prod – target deployment.
  - Sec – sink for data generated by Config Service and CloudTrail to S3 buckets.
- Encryption
  - KMS Encryption keys used throughout
  - Enforced SSL connections to Redshift
  - S3 – enforced write encryption (by policy).
- Audit and compliance documentation
  - AWS Artifacts.



Redshift topology

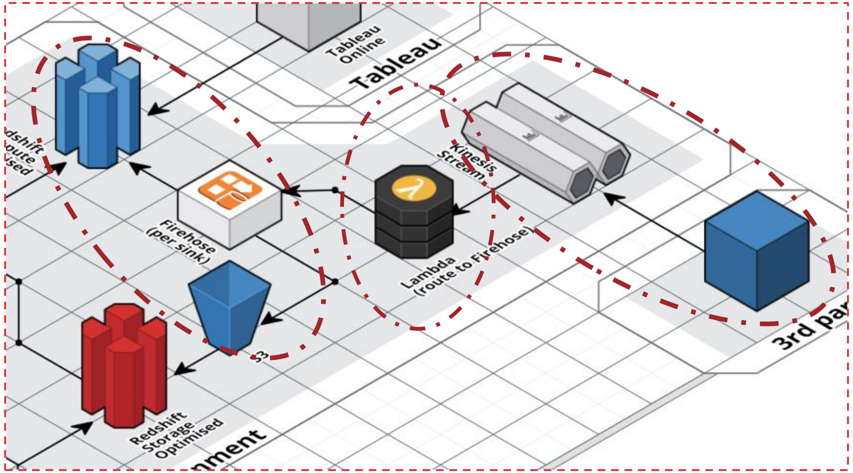
- Storage Optimised (red)
  - Optimised for storing larger volumes of data (source).
  - Ingestion point for newly arriving data.
  - Transformation layer (large number of work tables).
  - VPC - private subnet.

- Compute Optimised layer (blue)
  - Transformed data.
  - Near real-time operational data.
  - Present dimensional layer.
  - VPC – public subnet (whitelisted access).



What about Streaming?

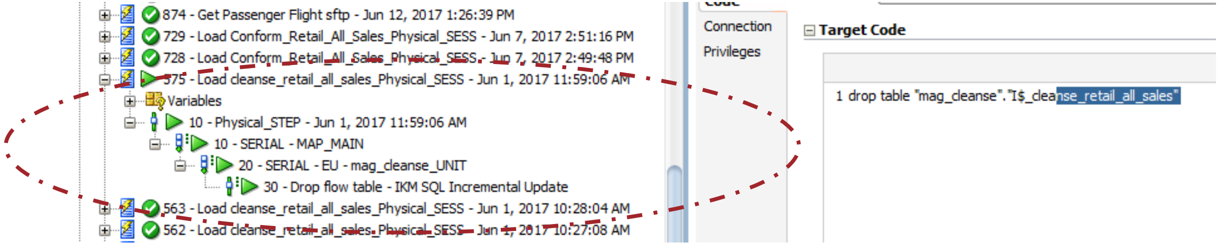
- Setup Kinesis Streams to allow 3rd parties to send data.
- Enabled cross account access with an assumed role.
- Used Lambda to route mixed data to multiple Firehose Streams.
- Firehose Streams sink data to S3 and/or Redshift Compute (blue).



Observations

ODI and Redshift

- Problem: ODI initiated Redshift tasks not completing.



```
874 - Get Passenger Flight sftp - Jun 12, 2017 1:26:39 PM
729 - Load Conform_Retail_All_Sales_Physical_SESS - Jun 7, 2017 2:51:16 PM
728 - Load Conform_Retail_All_Sales_Physical_SESS - Jun 7, 2017 2:49:48 PM
575 - Load cleanse_retail_all_sales_Physical_SESS - Jun 1, 2017 11:59:06 AM
  Variables
  10 - Physical_STEP - Jun 1, 2017 11:59:06 AM
    10 - SERIAL - MAP_MAIN
      20 - SERIAL - EU - mag_cleanse_UNIT
        30 - Drop flow table - IKM SQL Incremental Update
563 - Load cleanse_retail_all_sales_Physical_SESS - Jun 1, 2017 10:28:04 AM
562 - Load cleanse_retail_all_sales_Physical_SESS - Jun 1, 2017 10:27:08 AM
```

```
drop table "mag_cleanse"."I$_cleanse_retail_all_sales"
```

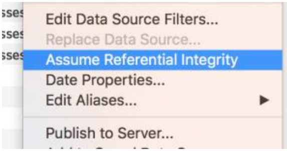
Solution: Increase Array Fetch Size in ODI

- Problem: No native knowledge modules in ODI for Redshift.
  - Solution: Customised existing generic SQL knowledge modules for Redshift.
    - Evaluating 3rd party solution Knowledge Module.

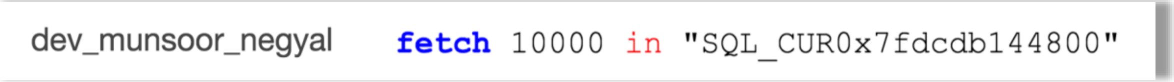
crimsonmacaw

Tableau and Redshift

- How does Tableau Online connect to Redshift?
  - JDBC via SSL.
  - Whitelisted to Redshift.
  - Tableau available in multiple regions (US, Ireland).
- Enable Redshift constraints:
  - Foreign Key and Primary Key and Unique constraints – ensure they are created in Redshift (even though they are not enforced).
- Enable Tableau "Assume Referential Integrity"
  - in Tableau workbooks (if you have it!).



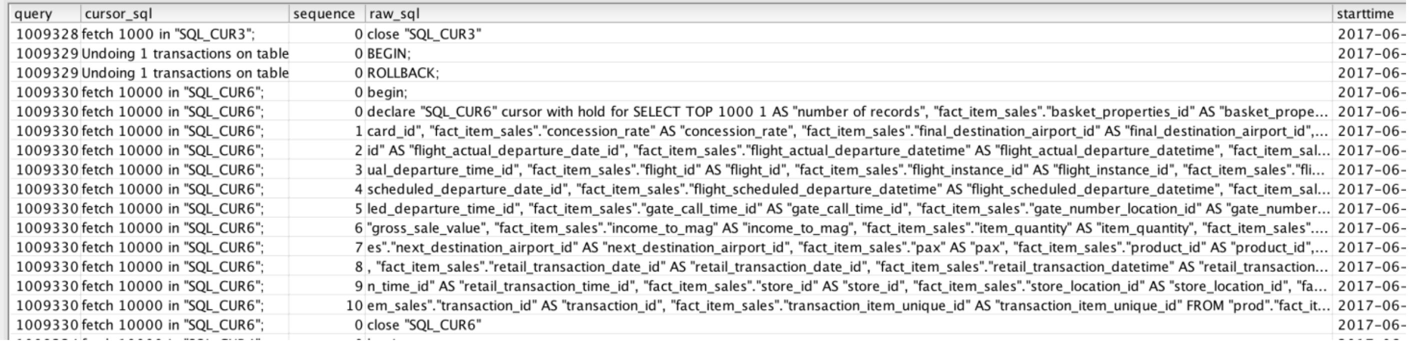
- Queries in Tableau:
  - Executed via Redshift cursor – minimise IO.
  - Current activity: `stv_active_cursors`.
  - For recent activity (two - five days): `stl_query` and `stl_utility_text`.



```
dev_munsoor_negyal    fetch 10000 in "SQL_CUR0x7fdcdb144800"
```

Tableau – getting back to SQL

```
select
         a.query,a.querytxt as cursor_sql
         ,b.sequence,b.text as raw_sql, b.starttime
from
         stl_query a inner join stl_utilitytext b
         on a.pid = b.pid and a.xid = b.xid
where
         database = ‘<DBName>’
         and a.starttime >= dateadd(day, -1, current_date)
order by
         a.xid, b.starttime, b.sequence asc;
```



query	cursor_sql	sequence	raw_sql	starttime
1009328	fetch 1000 in "SQL_CUR3";	0	close "SQL_CUR3"	2017-06-
1009329	Undoing 1 transactions on table	0	BEGIN;	2017-06-
1009329	Undoing 1 transactions on table	0	ROLLBACK;	2017-06-
1009330	fetch 10000 in "SQL_CUR6";	0	begin;	2017-06-
1009330	fetch 10000 in "SQL_CUR6";	0	declare "SQL_CUR6" cursor with hold for SELECT TOP 1000 1 AS "number of records", "fact_item_sales"."basket_properties_id" AS "basket_prope...	2017-06-
1009330	fetch 10000 in "SQL_CUR6";	1	card_id", "fact_item_sales"."concession_rate" AS "concession_rate", "fact_item_sales"."final_destination_airport_id" AS "final_destination_airport_id",...	2017-06-
1009330	fetch 10000 in "SQL_CUR6";	2	id" AS "flight_actual_departure_date_id", "fact_item_sales"."flight_actual_departure_datetime" AS "flight_actual_departure_datetime", "fact_item_sal...	2017-06-
1009330	fetch 10000 in "SQL_CUR6";	3	ual_departure_time_id", "fact_item_sales"."flight_id" AS "flight_id", "fact_item_sales"."flight_instance_id" AS "flight_instance_id", "fact_item_sales"."fli...	2017-06-
1009330	fetch 10000 in "SQL_CUR6";	4	scheduled_departure_date_id", "fact_item_sales"."flight_scheduled_departure_datetime" AS "flight_scheduled_departure_datetime", "fact_item_sal...	2017-06-
1009330	fetch 10000 in "SQL_CUR6";	5	led_departure_time_id", "fact_item_sales"."gate_call_time_id" AS "gate_call_time_id", "fact_item_sales"."gate_number_location_id" AS "gate_number...	2017-06-
1009330	fetch 10000 in "SQL_CUR6";	6	"gross_sale_value", "fact_item_sales"."income_to_mag" AS "income_to_mag", "fact_item_sales"."item_quantity" AS "item_quantity", "fact_item_sales"....	2017-06-
1009330	fetch 10000 in "SQL_CUR6";	7	es"."next_destination_airport_id" AS "next_destination_airport_id", "fact_item_sales"."pax" AS "pax", "fact_item_sales"."product_id" AS "product_id",...	2017-06-
1009330	fetch 10000 in "SQL_CUR6";	8	, "fact_item_sales"."retail_transaction_date_id" AS "retail_transaction_date_id", "fact_item_sales"."retail_transaction_datetime" AS "retail_transaction...	2017-06-
1009330	fetch 10000 in "SQL_CUR6";	9	n_time_id" AS "retail_transaction_time_id", "fact_item_sales"."store_id" AS "store_id", "fact_item_sales"."store_location_id" AS "store_location_id", "fa...	2017-06-
1009330	fetch 10000 in "SQL_CUR6";	10	em_sales"."transaction_id" AS "transaction_id", "fact_item_sales"."transaction_item_unique_id" AS "transaction_item_unique_id" FROM "prod"."fact_it...	2017-06-
1009330	fetch 10000 in "SQL_CUR6";	0	close "SQL_CUR6"	2017-06-

Redshift

- Performance so far has been very good.
- A lot to do with the design of Redshift.
    - Optimisations so far have been limited to:
        - Fields:
            - lengths
            - datatypes
            - compression datatypes.
        - Distribution keys.
        - Sort keys.
        - Skew analysis.
        - Vacuum and ANALYZE.
- But we intend to do some more work on below:
    - Work queue management.
    - User load analysis.
    - Attribute pushdown.

Fast

Cost Efficient

Simple

Elastic

Secure

Compatible

1. Analyze Database **Audit Logs** for Security and Compliance Using Amazon Redshift Spectrum amzn.to/2szR3nf
2. Build a **Healthcare Data Warehouse** Using Amazon EMR, Amazon Redshift, AWS Lambda, and OMOP amzn.to/2rr7LWq
3. Run **Mixed Workloads** with Amazon Redshift Workload Management amzn.to/2srlL1g
4. **Converging Data Silos** to Amazon Redshift Using AWS **DMS** amzn.to/2kIr1bq
5. Powering Amazon Redshift Analytics with Apache **Spark** and **Amazon Machine Learning** amzn.to/2rgR8Z7
6. Using pgpool and Amazon ElastiCache for **Query Caching** with Amazon Redshift amzn.to/2Ir66MH
7. Extending Seven Bridges Genomics with Amazon **Redshift and R** amzn.to/2tlylga
8. **Zero Admin** Lambda based Redshift Loader bit.ly/2swvvl6

Thank You

Data is magic!

amazon
web services