snowflake

CLOUD ANALYTICS CONFERENCE

LONDON

@SnowflakeDB #CloudAnalytics17

This Afternoon….

- 5 Ways to Enable BI in the Cloud – (High level summary)
- Optimizing Your Analytics with Tableau and Snowflake – (Detailed BI content)
- Utilizing Snowflake's Architecture to Support BI – (Detailed Snowflake content)
- Start Ending Your Data Struggle (30 Day Guide)

Struggle to Analyze

5 Ways to Enable BI in the Cloud

Erika Bakse, IAC

Ross Perez, Snowflake

6/1/2017

Agenda - 5 Ways to Enable BI in the Cloud

- What does the BI team need to accomplish?
  - Key objectives for the business intelligence team
- Why is it so hard to get there?
  - 5 analytics roadblocks and how to solve them
- How did we Enable BI in the "Real World" at IAC?

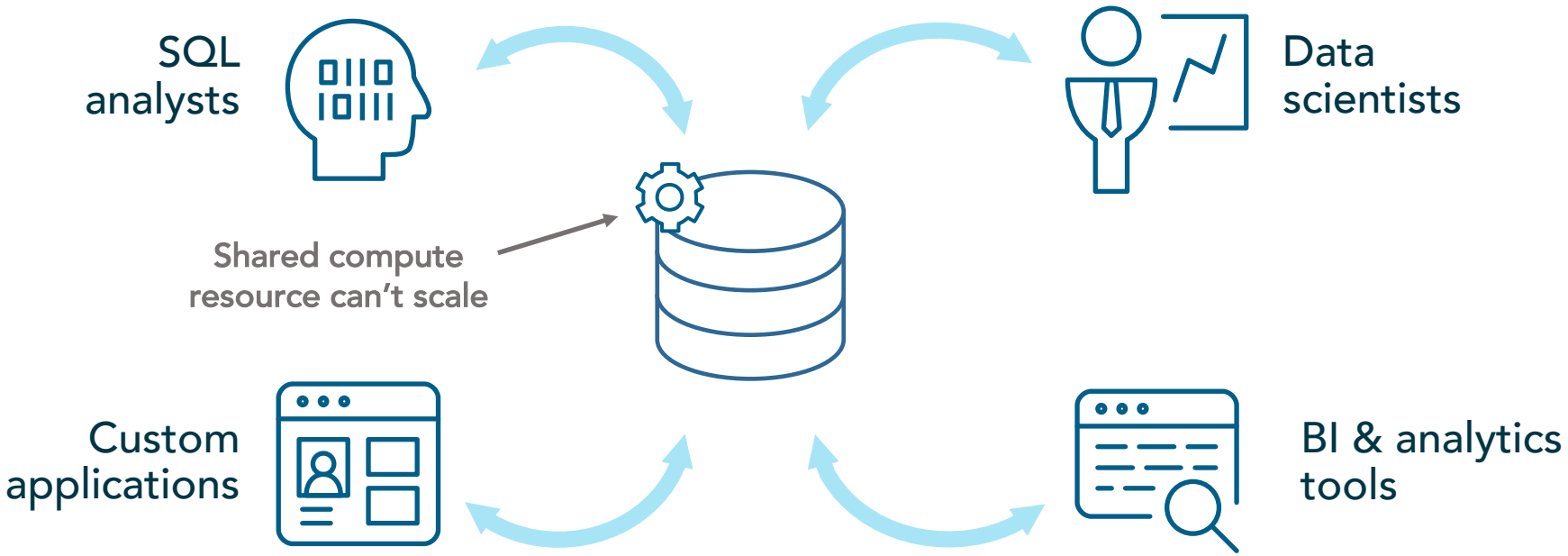
Key objectives for the business intelligence team

- Deliver information and insight to the organization
  - Enable data access for everyone with a need
  - Maintain accurate and relevant data
  - Keep overhead light to focus on analytics

Common analytics problems (and how to solve them)

- Concurrency
- Complexity
- Scalability
- Access to data
- Cost

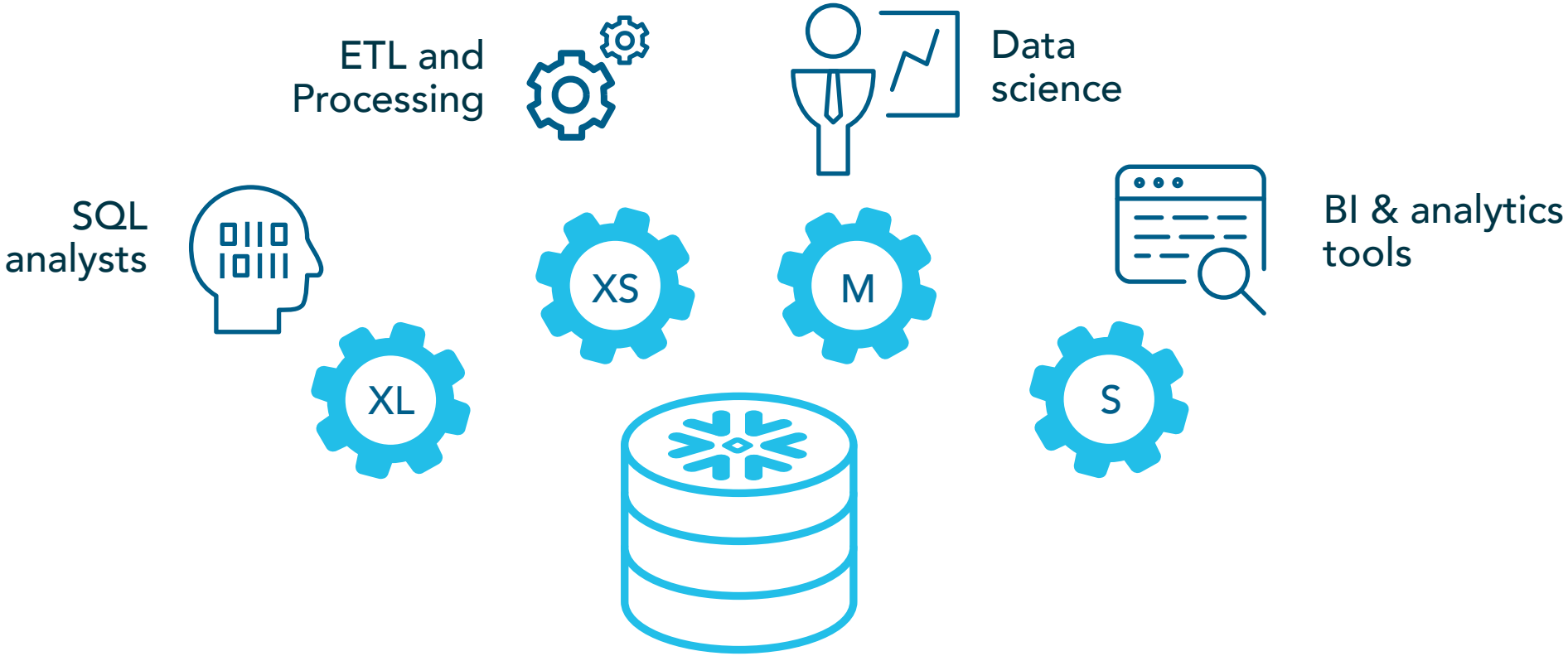
1. Concurrency



1. Concurrency “solutions”

- Data marts
  - Manual effort and syncing
- Extracts and in-memory solutions
  - Curation, staleness and connectivity
- Making a copy of the database
  - Takes time and effort, increases cost
- Scheduled or restricted access
  - Reduces the value of the data to the team

1. Using the cloud to eliminate concurrency



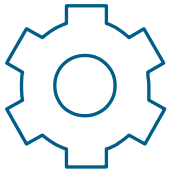
2. Complexity

- Manual tuning and optimization
- Tuning in one area leads to degradation in another
- In smaller companies, sometimes complex tools aren't optimized at all

2. Complexity “solutions”

- A dedicated resource is hired to manage the database
  - Expensive and time consuming
- A decision is made about which queries to tune for
  - Only the “tuned” queries are performant
- Analysts have to act as their own DBA
  - Inefficient use of resources

2. Addressing complexity with Snowflake

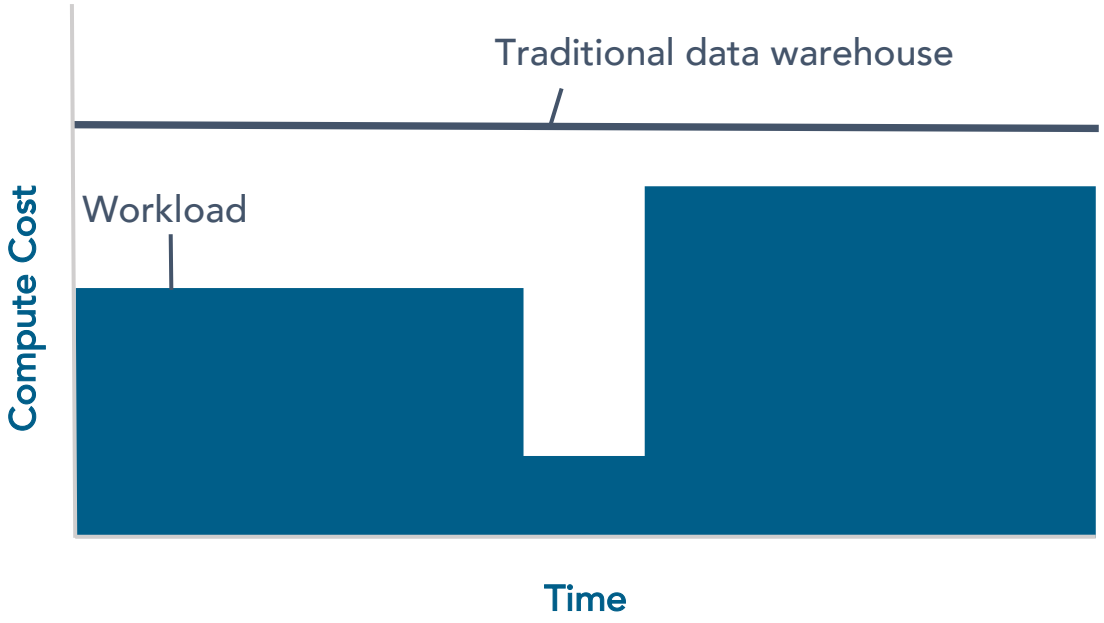


No knobs to turn or tune

Indexed and optimized automatically for all queries

Standard SQL

3. Scalability

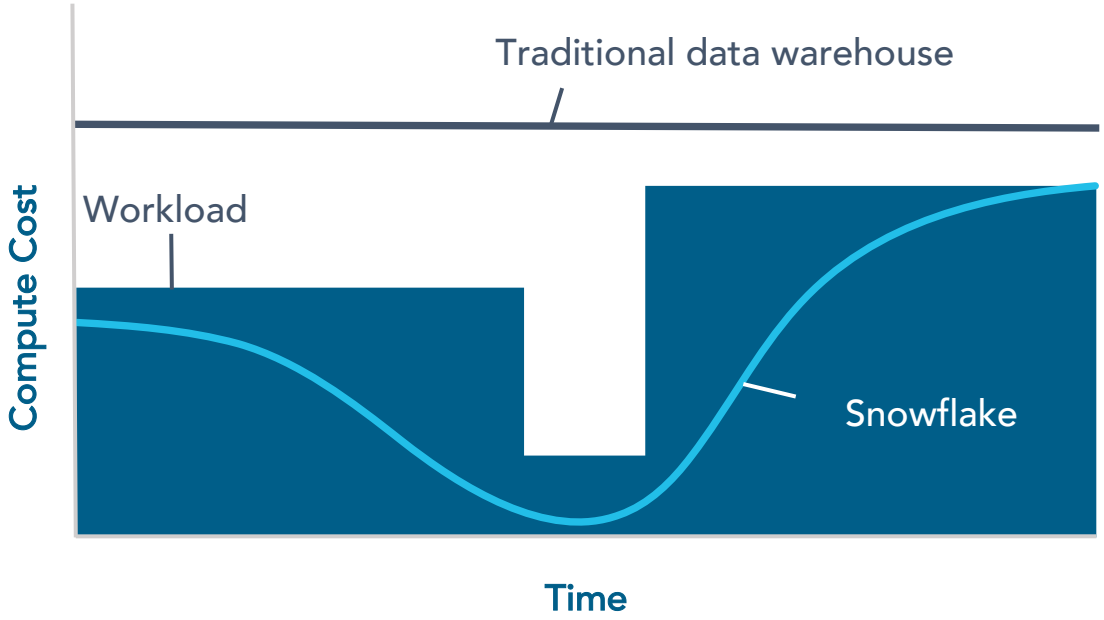


3. Scalability

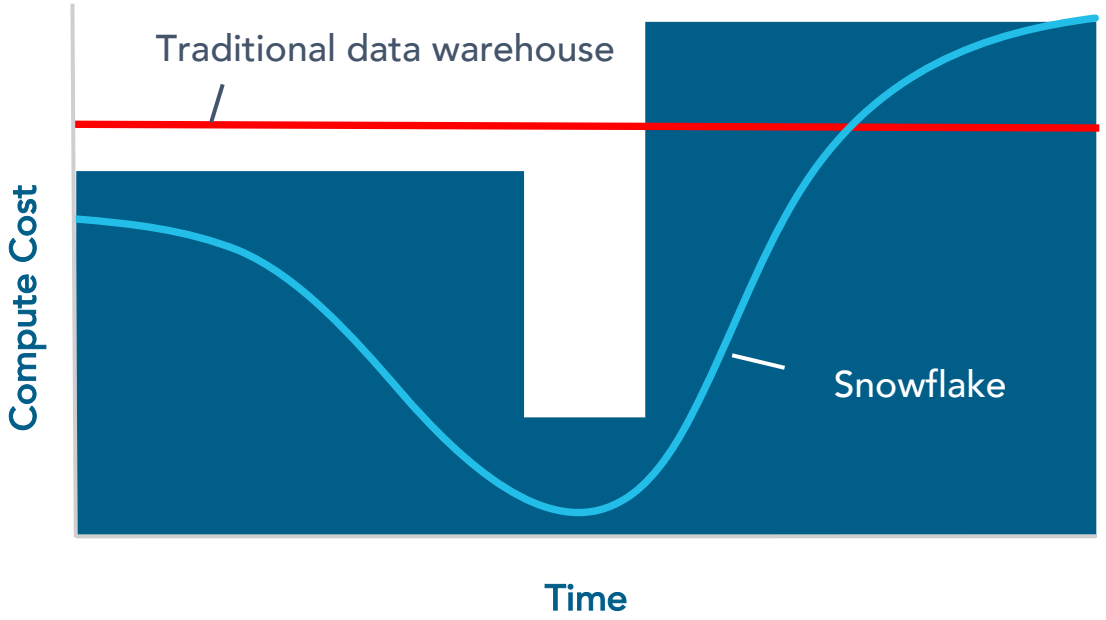
3. Scalability “solution”

- Buy more database!
  - Expensive and marginally inefficient, requires shut down and restart

3. Snowflake uses the cloud to match demand



3. Snowflake uses the cloud to match demand



4. Access to data

- Database only available at certain times by certain people
- Database complexity reduces ability for analysts to effectively use
- Local access only

4. Access to data “solutions”

- Request queue
  - Slow, inefficient and frustrating
- Create curated data sources
  - Requires maintenance and updates
- VPN
  - Inefficient and management intensive

4. Addressing access to data with Snowflake

- Virtual warehouses enable dedicated resources for each use case
- Standard SQL mitigates complexity problems
- Cloud access means simplified access

5. Cost

- Every single one of these problems is solved partially with
  - More money
  - More people (so therefore more money)
  - More time (and therefore more people and more money)

5. Addressing cost with Snowflake

- Pay for what you use
- Scale up and down however necessary
- No up front commitment

The IAC Publishing Labs Story

You may know us AskJeeves! …but we do more than just that:

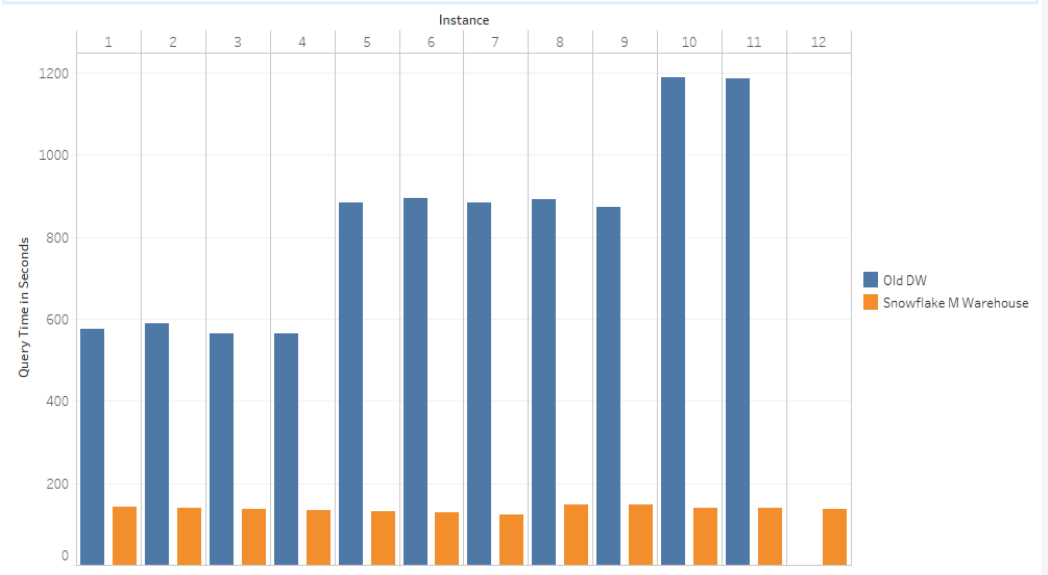
- 300+ million events and 50-100 million keywords
- Manage marketing terms for bidding and monetization
- Import more than 1.5 terabytes of raw data daily
- Multiple data sources (Google Adsense, Bing, Yahoo, Buy-side cost data, etc)

Our Analytics Problems Sound Familiar…

- Concurrency
  - Need to process big data, but fighting with business users
  - Rogue queries (or a misclick…) bring down the entire cluster
  - No separated test/dev environment!
- Complexity
  - No dedicated on-site data engineer
  - Have to understand distribution keys, sort keys, compression, oh my!
  - Can you say VACCUUM?
- Scalability
  - We had a 26 node cluster, yet could only support 6 months of raw data
  - What about new initiatives? Expanded SEM, Data Science, Content….
- Access to data
  - Concurrency issues limit access
  - JSON blobs require heavy processing
  - Separate Hadoop cluster to handle raw data outside of retention limits
- Cost

IACPL: Concurrency

- This was our **number one** biggest issue
  - Need to process big data, but fighting with business users
    - With Snowflake, we process most of our data on an XS Warehouse which runs 24x7
    - Result: **Cleaner Data Faster Cheaper**
  - Rogue queries (or a misclick…) bring down the entire cluster
    - Separate L Warehouse during business hours for analysts
    - Ability to separate out business groups by warehouse
    - Easy to spin up separate adhoc warehouses for intensive queries
  - No separated test/dev environment!
    - Snowflake CLONE quickly and easily gives us a test environment



IACPL: Complexity

- Have to understand distribution keys, sort keys, compression, oh my!
  - Not necessary for Snowflake
  - Specify clustering keys if you need to tune, check with CLUSTERING_RATIO
    - I tested this on a slow query. Results? Snowflake's clustering was better than mine. The issue was a WHERE clause!
- Can you say VACCUUM?
  - Not anymore!

IACPL: Scalability

- We had a 26 node cluster, yet we could only retain 6 months of raw data before hitting space limits
  - Before migration we had 65 TB, post migration we had 35TB compressed
  - Now maintain 18 months raw data, plus expanded offerings and we are only at about 55TB
- What about new initiatives? Expanded SEM, Data Science, Content….
  - We are now in the business of saying YES
  - Storage costs in Snowflake so affordable we just throw it all up there: fantastic for momentum on POC efforts

IACPL: Access To Data

- Concurrency issues already limiting user access to certain times
- JSON blobs require heavy processing
  - Native JSON querying means no need to alter DDLs or do heavy engineering when changes are made
  - Increased flexibility and quicker determinations of value
- Separate Hadoop cluster to handle raw data outside of retention limits
  - Now maintain 18 months of raw data in Snowflake (the legal limit without anonymization/rollups)
  - Data Science team no longer expending effort to duplicate our work structuring data
  - Hadoop cluster has been deprecated – but if necessary we'll spin it up on AWS

IACPL: Cost

- We saved **78%** of our data warehousing costs migrating to Snowflake vs upgrading our on-premises product

Thank You to Our Partners

Platinum

amazon web services

tableau

talend

tdwi
THE DATA WAREHOUSING INSTITUTE

Gold

looker

alteryx

bigdata INTERNATIONAL

Cervello

snapLogic

Information Builders

snowflake
CLOUD ANALYTICS
CONFERENCE