Implementing a new database platform for your modern data warehouse – Snowflake on Azure

Grega Jerkič
(grega.jerkic@in516ht.com)

CEO, IN516HT D.O.O.

know your numbers

Agenda

- About me
- Snowflake story
  - Elasticity & “Unlimited” Concurrency
  - Separation of compute & storage
  - Powerful support for semi-structured data
  - Data Sharehouse – simplifying data exchange and new business ideas
- Use cases and integration with Microsoft Azure stack

insight
know your numbers

About Me

- Working with Microsoft SQL server from 1999.
- Data developer, consultant and mentor with a focus on the implementation of complex analytical systems.
- Training activities around Europe for Microsoft and IBM.
- Snowflake person of the year 2018 for Central Europe.
- Co-Author for Training Kit (Exam 70-463) Implementing a Data Warehouse with Microsoft SQL Server 2012
- Managing director and co-owner of the company In516ht.
- Developed WW solution on top of Microsoft Dynamics (in 2001 – 2006) – www.bi4dynamics.com

IN516HT [insight]
Know Your Numbers

Helping Companies Be Data Driven

HQ in Slovenia

Focus:
- CEE region
- Middle East

Team with 15+ years experience building complex data warehouses.

Data Warehouse Modernization
Cloud data warehouse, Big Data, IOT, data integration, IBM Industry Models (banking, insurance, retail).

Data Science
Predictive analytics, AI and optimization problems.

Deep Business Experience and Know-How
Banking, Insurance, Retail, Energy, Pharma and Telecommunication:
- Largest insurance group in the region – Triglav Group
- Large banking group in the region – NLB Group
- Retail & Energy – 30+ companies – Petrol Group

Snowflake Story

SQL Data Warehouse built for the cloud

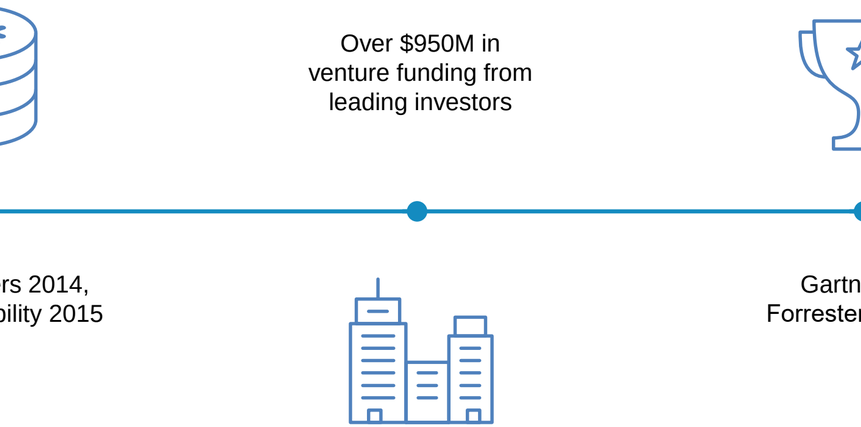
Founded 2012
by industry
veterans

First customers 2014,
general availability 2015

Over $950M in
venture funding from
leading investors

Gartner and
Forrester “Leader”

Over 1,500
active customers



Queries processed in Snowflake per day:	**> 60.000.000**
# rows in largest single table:	**68.000.000.000.000**
Largest number of tables single DB:	**200,000**
Single customer most data:	**> 40 PB**
Single customer most users:	**> 10,000**

insight
know your numbers

From 0 to 12 to … locations

- First office in EMEA in March 2017
- Over 150 customers
  in 12 countries
- 75 technology and solution partners

insight
know your numbers

How we decided to become a partner?

- Working from 2005 on different analytical appliances (MPP, columnar store) – simplicity & performance ware main drivers
- In summer 2017 we decided to test different cloud solutions since we had a lot of questions from our clients regarding Data Lake architectures
- Azure SQL DW, Redshift, Vertica, MemSQL, different GPU databases, Google BigQuery and Snowflake
- We took the data warehouse from one of our most complex clients (2-level data warehouse, atomic + dimensional, 1500+ ETL jobs, running on full rack of IBM Netezza)

It just worked

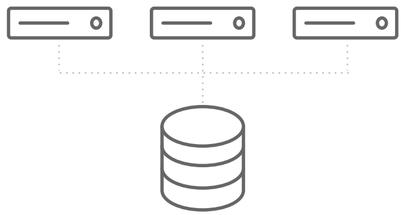
- We tested loading of data, needed changes to SQL, performance of most complex SQL insert jobs (1B+ fact with 30+ left joins to 100M tables) and different BI queries, scalability and administration (simplicity that we ware used)
- Key facts for selecting Snowflake:
  - Easy to migrate existing SQL procedures - Full SQL ANSI support (also all window functions) + some great additional extensions
  - Unique parallel execution for “unlimited” scalability
  - Performance at the friction of cost compared to large appliances
  - No administration and almost no tuning requirements
  - Second time billing and affordable prices – useful for small companies and huge enterprises
  - SaaS solution (cloud agnostic) – not PaaS or some hybrid

Key Snowflake proposition

New architecture for data warehousing

Multi-cluster, shared data, in the cloud

Traditional Architectures



Shared-disk

Shared storage
Single cluster

Shared-nothing

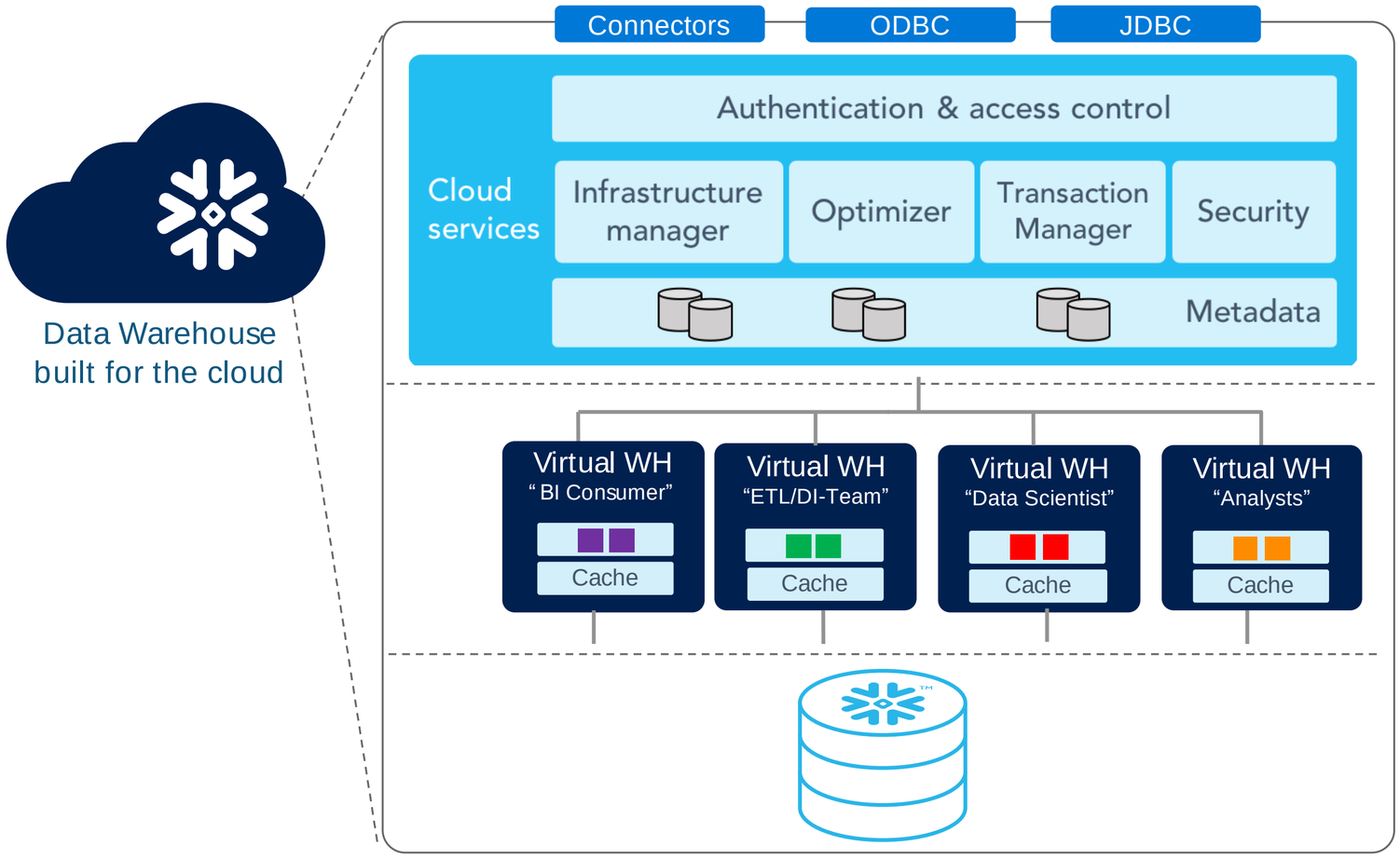
Decentralized,
local storage
Single cluster

Snowflake

Multi-cluster, shared data

- Centralized, scale-out storage
- Multiple, independent compute clusters

How Snowflake works



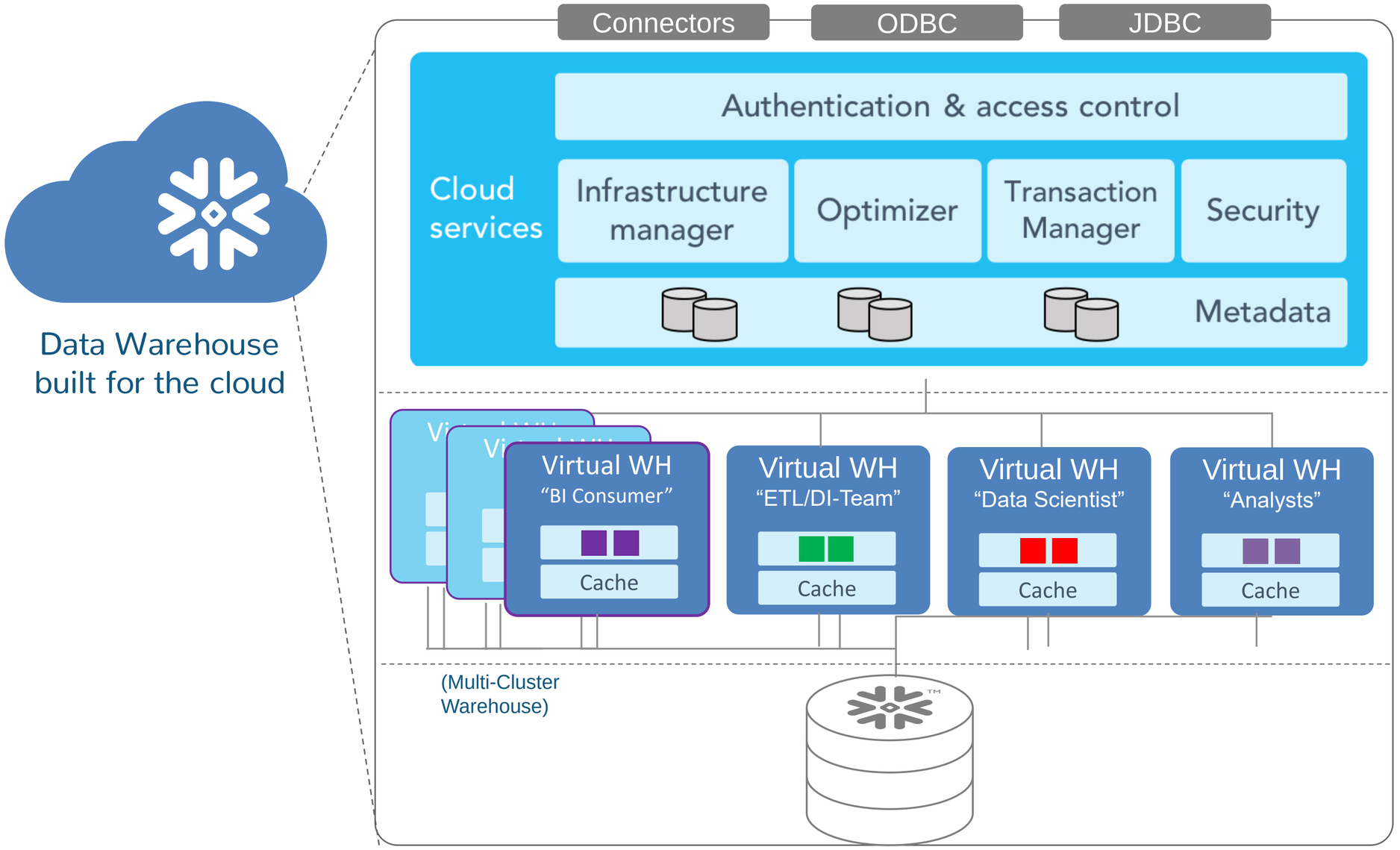
- Cloud Services
- Independently Scalable Compute
- Centralized Storage

insight
know your numbers

Elasticity & “Unlimited” Concurrency

DEMO

Elastic compute – for high concurrency



SCENARIO #1

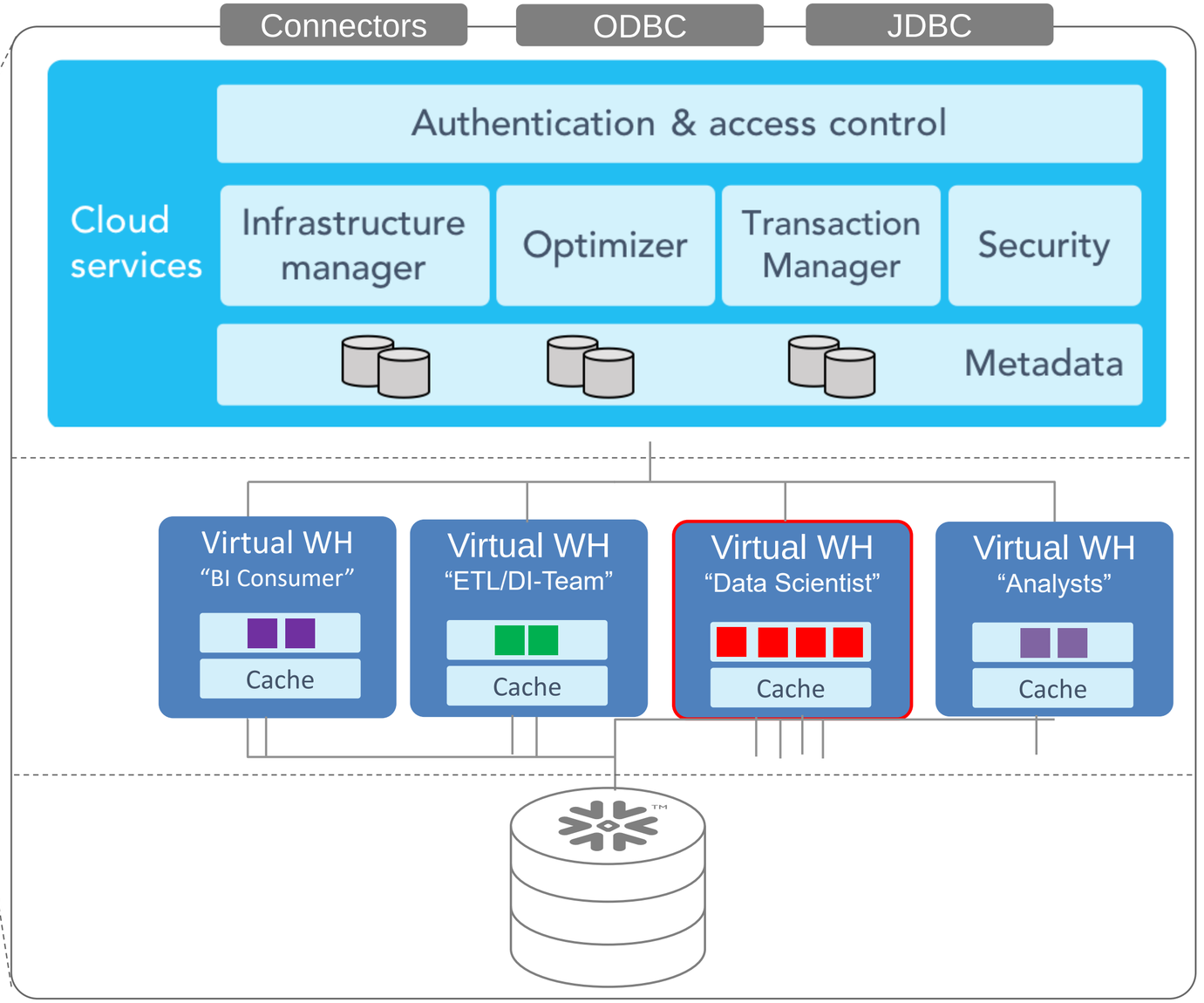
BI Consumer

- 10x more **concurrent user** sessions during reporting peak time, i.e. Monday morning dashboard calls, end of month reporting, etc.
- Snowflake **Multi-Cluster Virtual WH** automatically detects when to start another cluster and distributes all incoming queries via load balancer

insight
know your numbers

Elastic compute – for faster queries

Data Warehouse
built for the cloud



SCENARIO #2

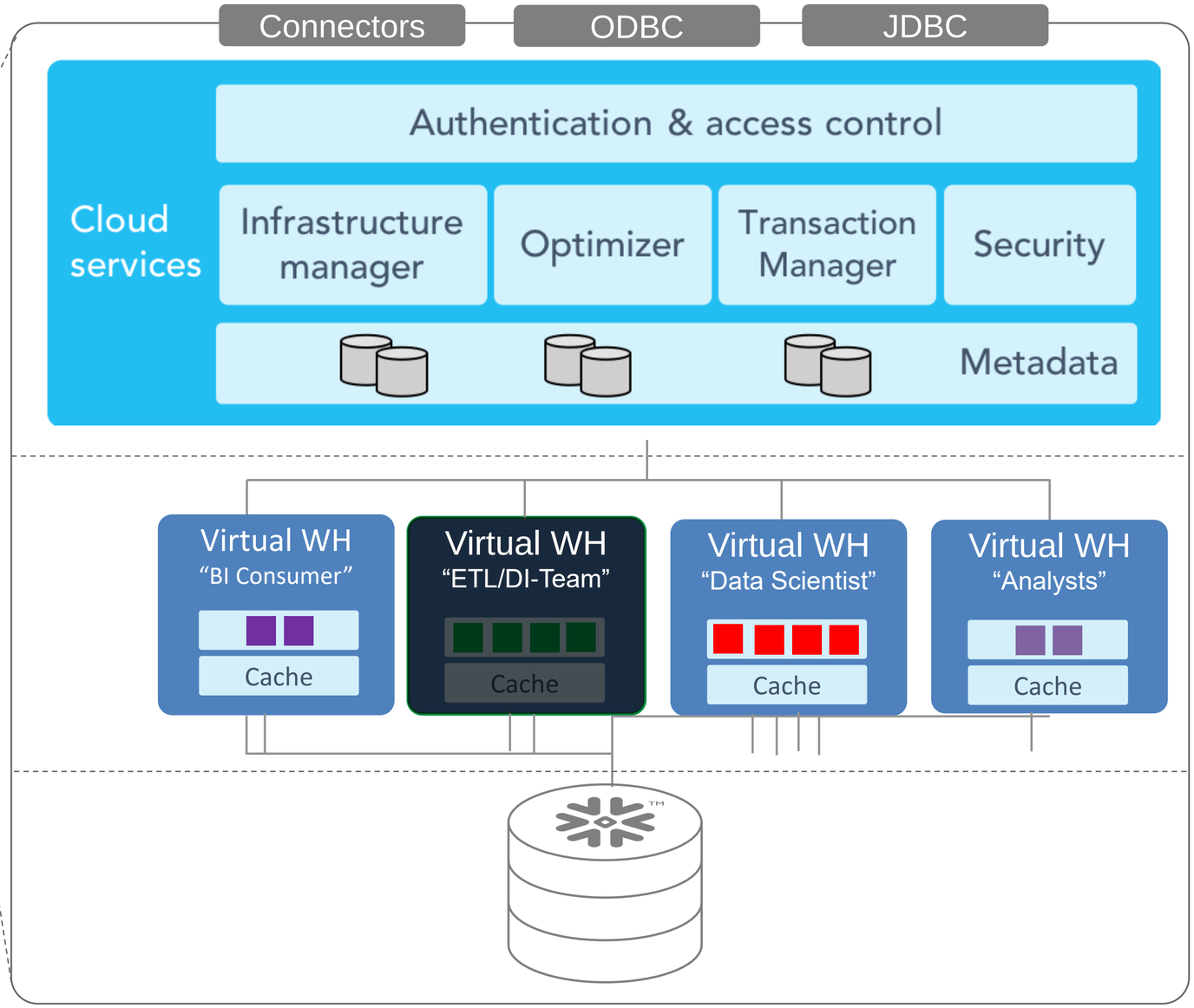
Data Scientist Team

- needs more compute resources for data preparation required for a new sales forecast, KPI
- is **instantly scaling-up** their **Snowflake Virtual WH** size to cut down query execution time

insight
know your numbers

Elastic compute – for faster data loads

Data Warehouse built for the cloud



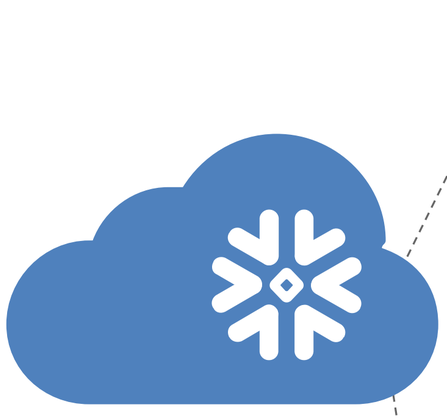
SCENARIO #3

ETL/DI-Team

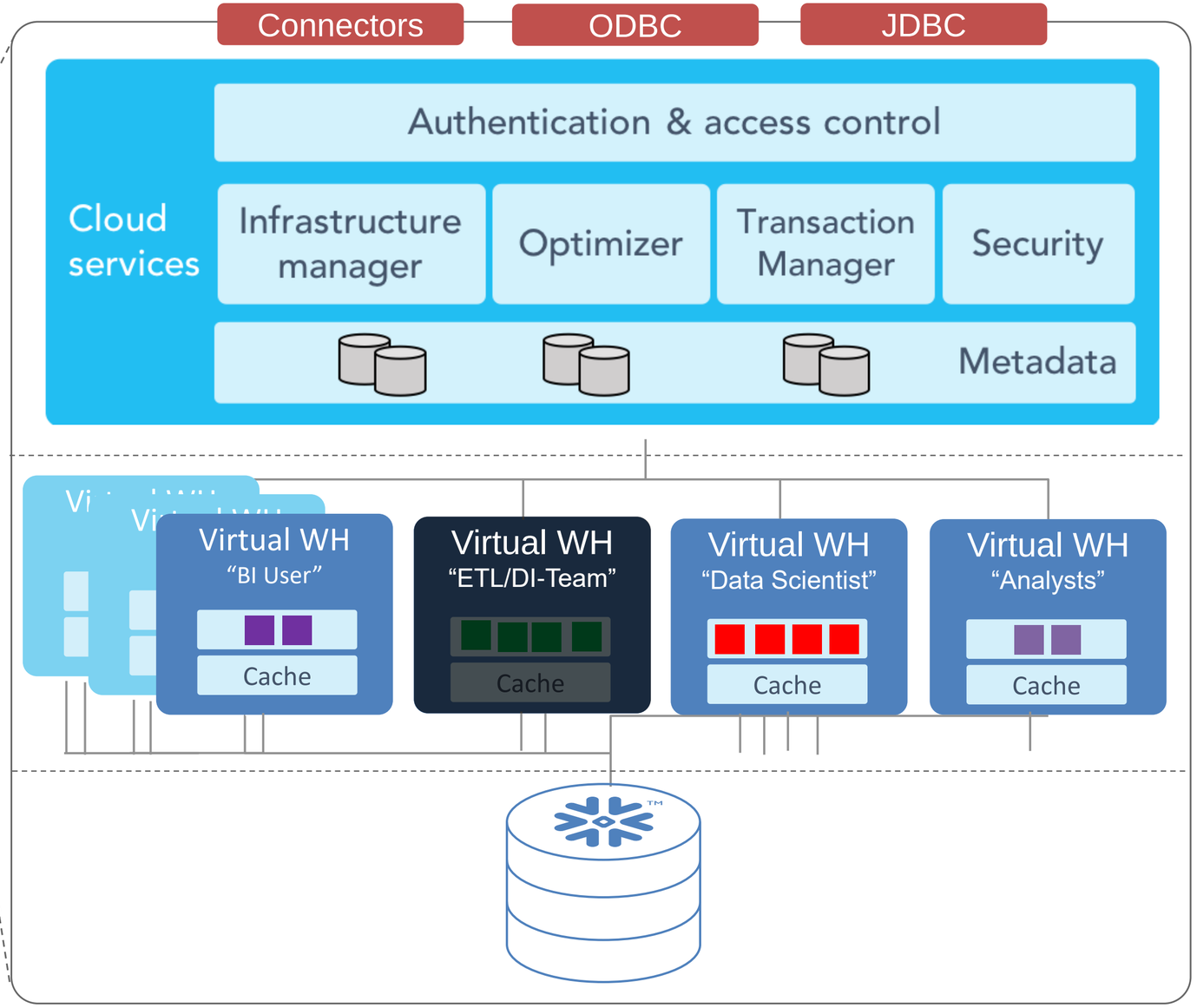
- needs more compute resources to complete a monthly data load in time
- uses a scheduler to **scale-up** their Snowflake **Virtual WH via SQL commands** orchestrated in a script
- once the data load is complete, the entire **Virtual WH suspends automatically**

insight know your numbers

Multi-cluster, shared data architecture



Data Warehouse
built for the cloud



Cloud Services

Scalable, resilient cloud services layer coordinates access & management

Independently Scalable Compute

Multiple "**Virtual Warehouses**" compute clusters scale horsepower & concurrency

Centralized Storage

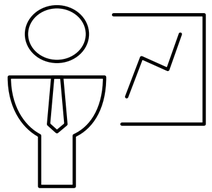
Instant, automatic scalability & elasticity

insight
know your numbers

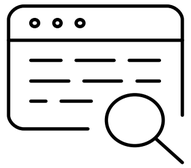
Pay for what you use…down to the second

ETL and Processing

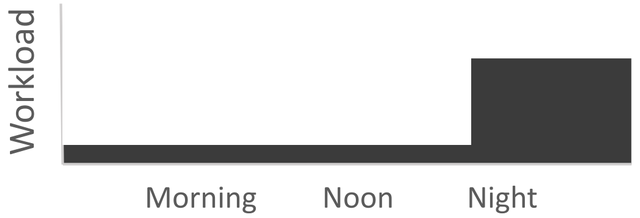
Reporting

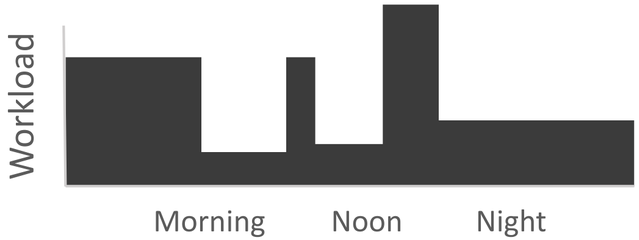


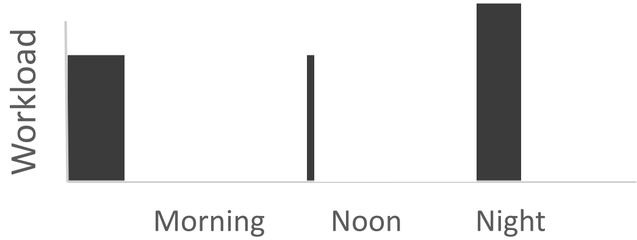
Ad-hoc Analytics

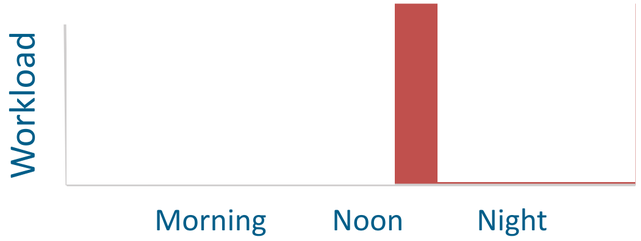


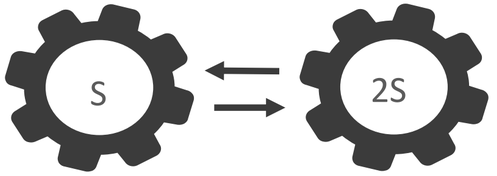
Data Scientist



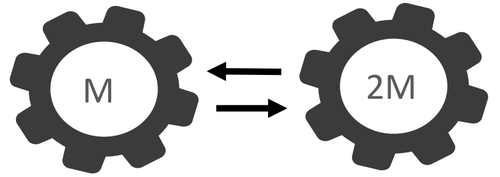




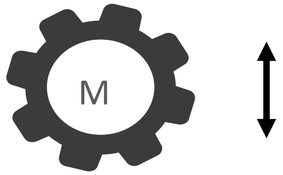




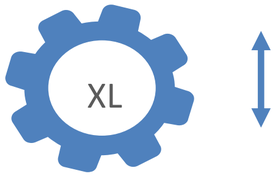
Autoscaling Multi-cluster Warehouse



Autoscaling Multi-cluster Warehouse



Scale up/down,
Auto-Suspend/Auto-Resume

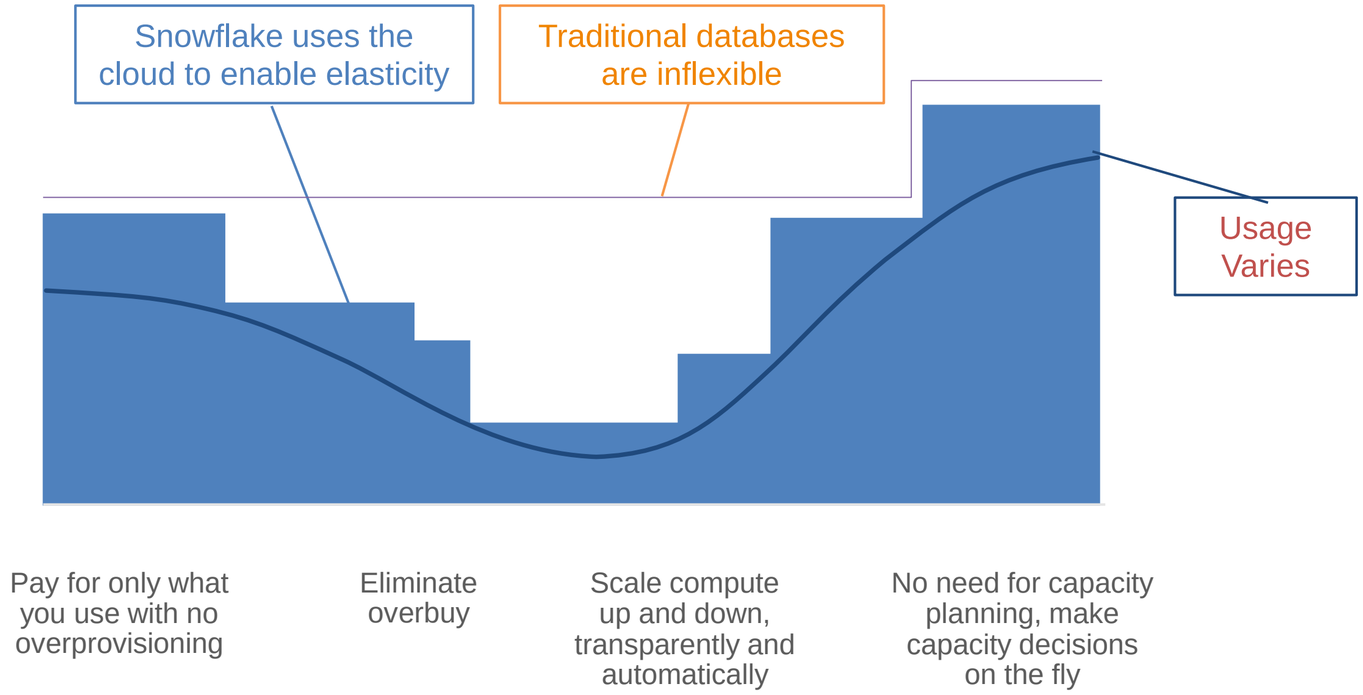


Scale up/down,
Auto-Suspend/Auto-Resume

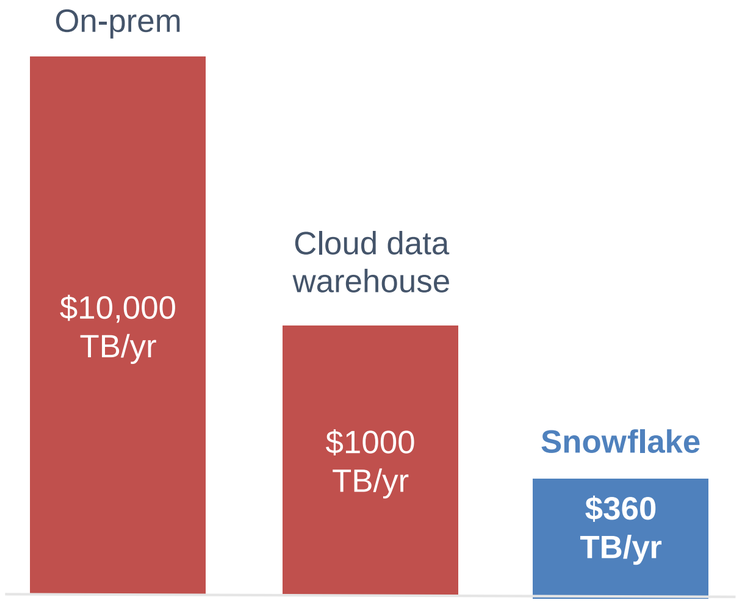
insight
know your numbers

Fast elasticity

Pay for only what you use, at cloud economies of scale



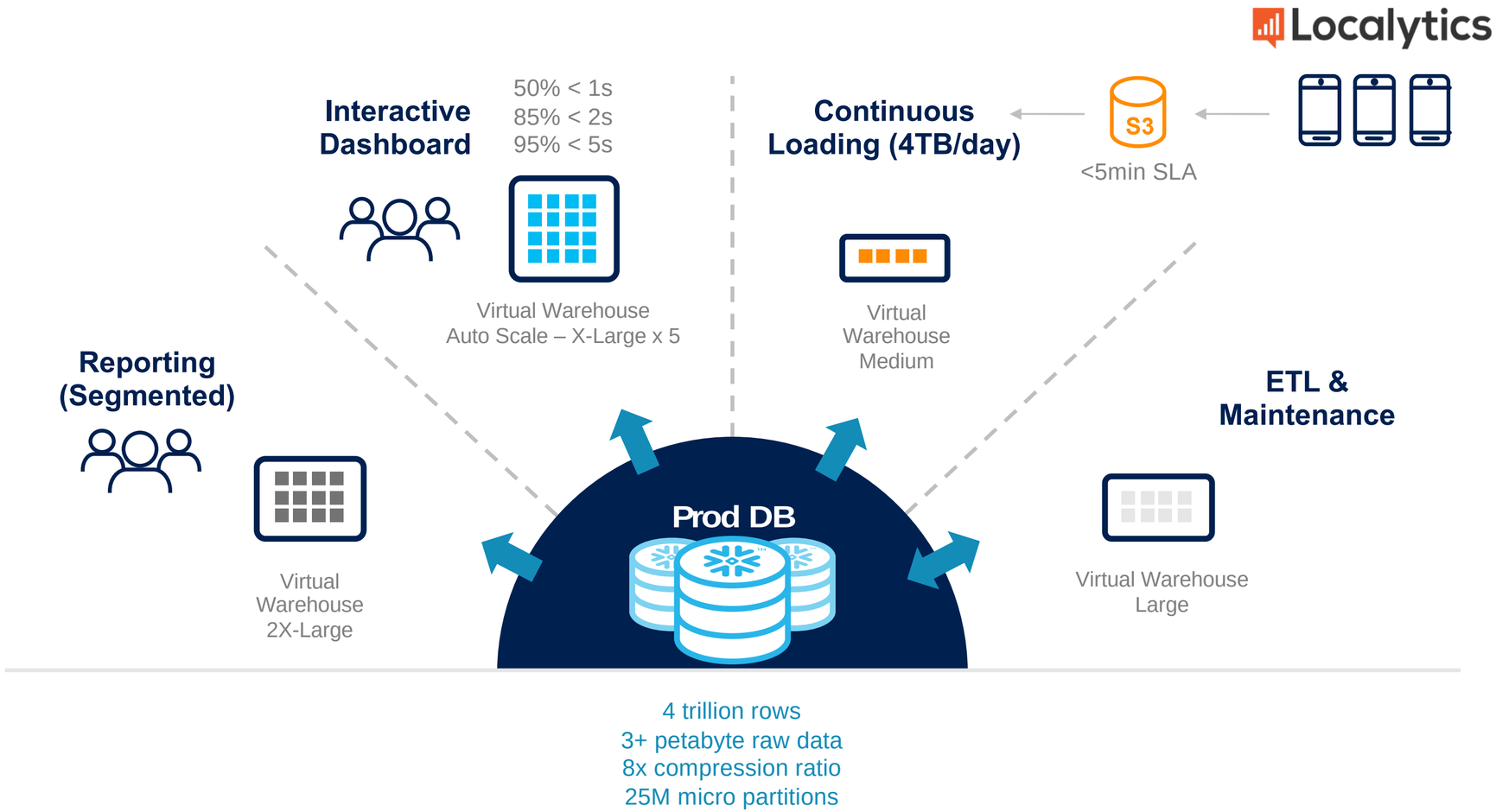
Pay only for what you use



Cloud economies of scale

insight
know your numbers

Real world use cases



Separation of compute & storage

DEMO

Back to the future

Time Travel (up-to 90 days)

insight
know your numbers

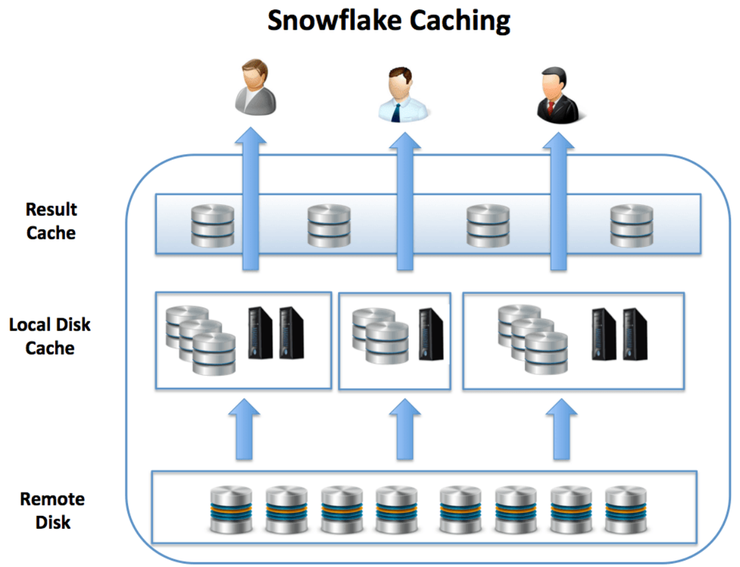
Cloning

- Amazing functionality
- Dev and Test environment
- DevOps practices can finally be implemented in the data intensive solutions
- Backup becomes irrelevant

insight
know your numbers

Strong Caching Mechanism

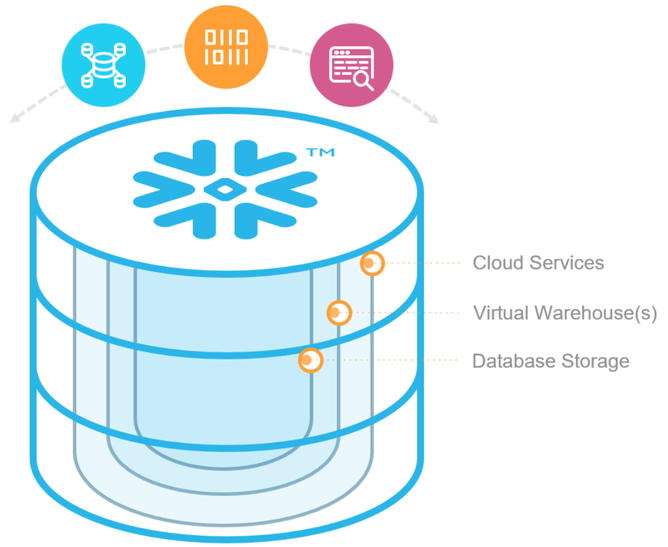
- No configuration
- ALL results persist for 24h
- Infinite space for caching (S3, Azure Blob)
- Cache available from all warehouse-s
- Fast results – no processing



Metadata
Metadata cached for fast access during query planning

Data
Active working set transparently cached on virtual warehouse SSD

Query results
Results sets cached for reuse without requiring compute
(e.g., static dashboard queries)

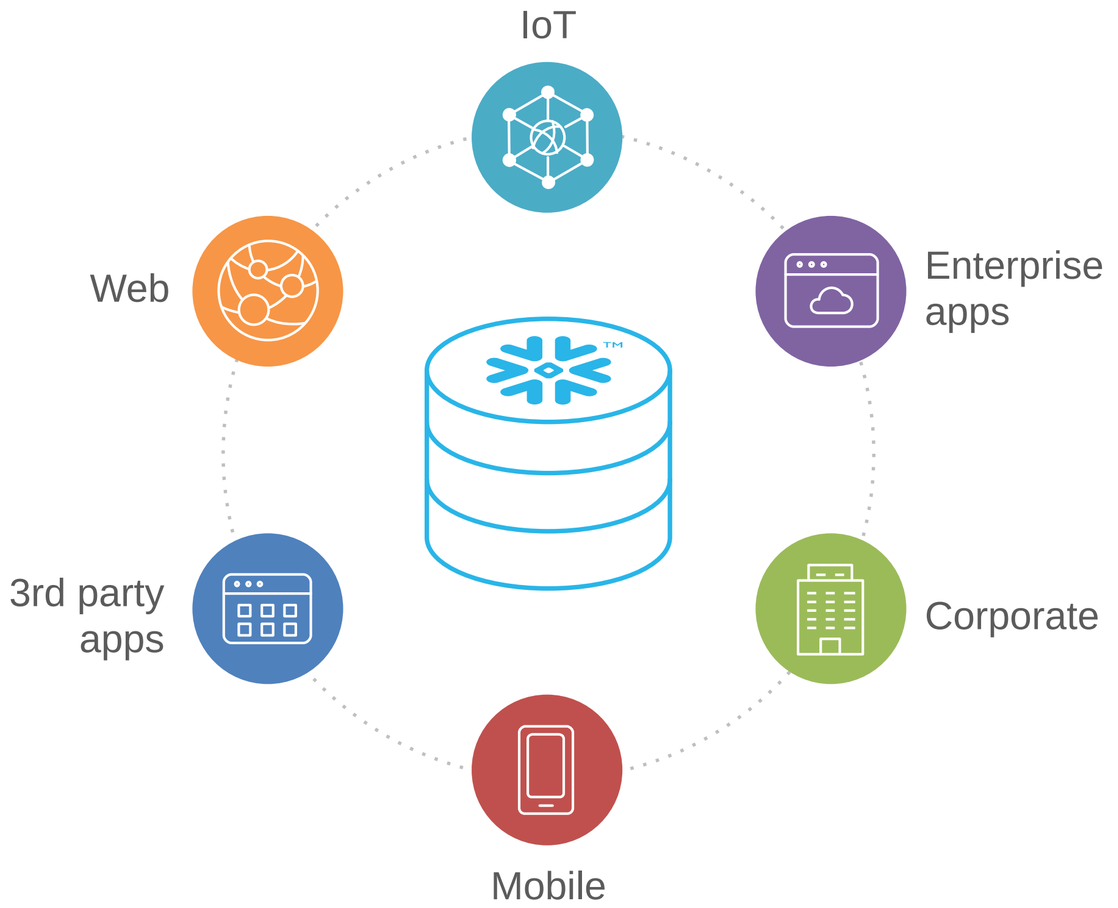


Powerful support for semi-structured data

DEMO

Powerful support for semi-structured data

- JSON
  - Built-in fast parser
  - Seamless integration with SQL and relational data
- XML
- AVRO
- ORC
- Parquet



XML experience

- Loading external business data (KPIs) of different (50k+) European companies
- One large XML (3GB+) file

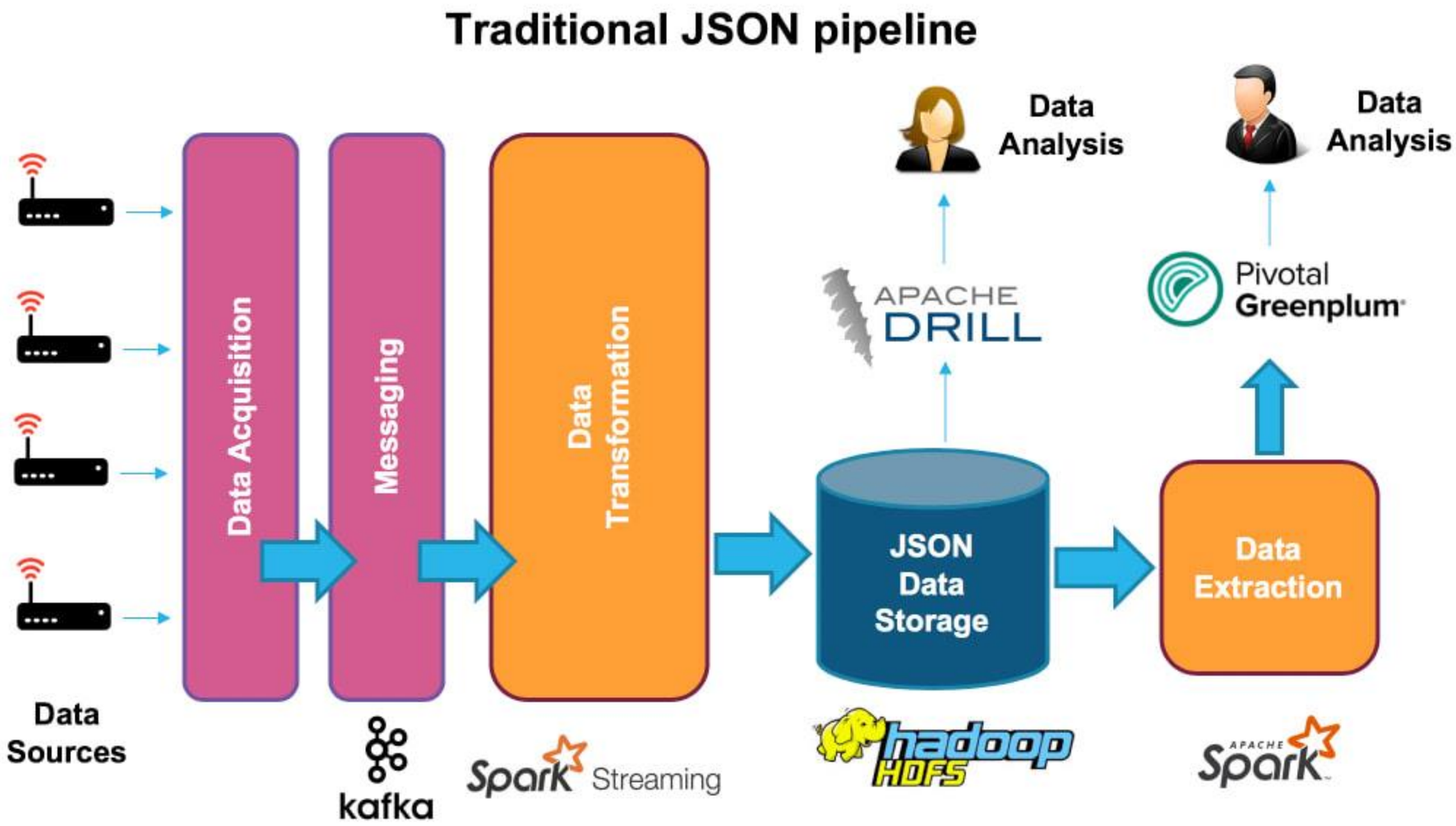
- Wrote our own XML splitter in to smaller chunks to be able to load in to the Snowflake (16 MB limitation)
- Elegant support for semi-structured data with materialized view support now

```
SELECT
    xxx.FILE_NAME AS filename,
    GET(xml_doc.value,'@id')::VARCHAR as ID,
    GET(items.value,'@field')::VARCHAR as fieldname,
    GET(items.value,'@fieldType')::VARCHAR as fieldtype,
    TRY_CAST(GET(items.value,'@index')::VARCHAR AS INTEGER) as SOME_VALUE,
    GET(items.value,'$')::VARCHAR as v
    FROM
        (
        SELECT FILE_CONTENT AS src_xml, FILE_NAME FROM DWH_DEV.APLLOAD.SEMI_EX_STORAGE WHERE FILE_CONTEXT = 'TEST_CONTEXT'
        )
    xxx,
    LATERAL FLATTEN(to_array($1:"$")) xml_doc,
    LATERAL FLATTEN(to_array(xml_doc.VALUE:"$")) items
```

insight
know your numbers

Simplification for data-driven applications

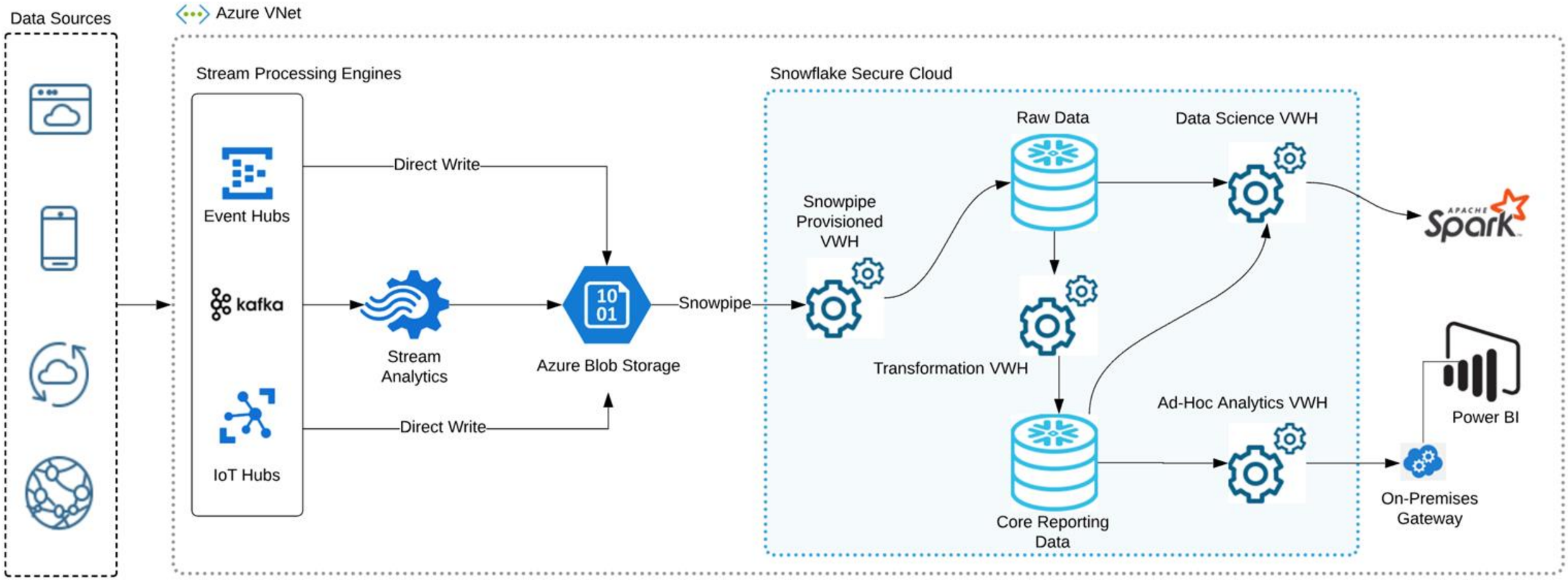
From Complex



insight
know your numbers

Simplification for data-driven applications

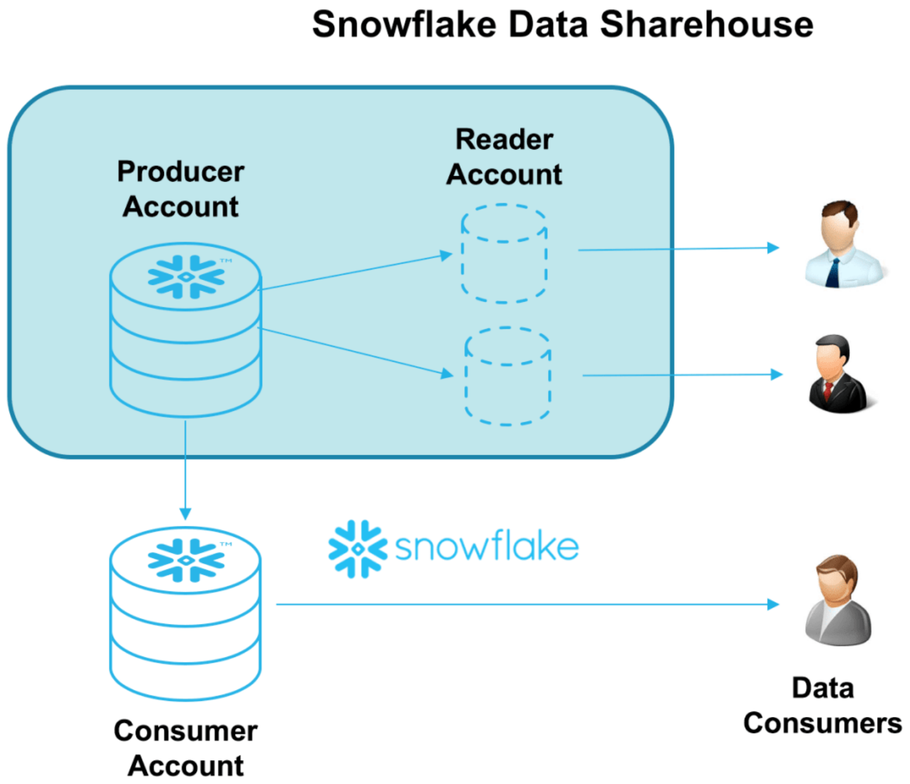
Prototype your solution in days



Data Sharehouse: simplifying data exchange and new business ideas

Data sharing and the new data economy

- Interesting and innovative concept
- Share your data as a normal folder
- No need for API development, just simple access to a table
- Data Economy cases are running Snowflake to sell their data
- Additional benefit for clients sharing their data sets



insight
know your numbers

How it works?

- A Producer creates one or more Data Shares to securely expose selected data to external clients
  - Existing Snowflake customers
  - Have the option of creating a read-only database which links directly to the Data Share. Using this method, consumers can directly execute queries against the shared data, and either enrich it in real time with other data sources, or analyse and query as required.
- Other customers
  - Can be provided with a Reader Account complete with the ability to connect the consumers preferred data analysis tool to the producer's data share.

Large cases in USA & UK

PLAYFAB

"With Snowflake Data Sharing, every PlayFab customer can have a world-class data warehouse, pre-populated with all their data, and updated in near real-time."

James Gwertzman, CEO and Co-Founder

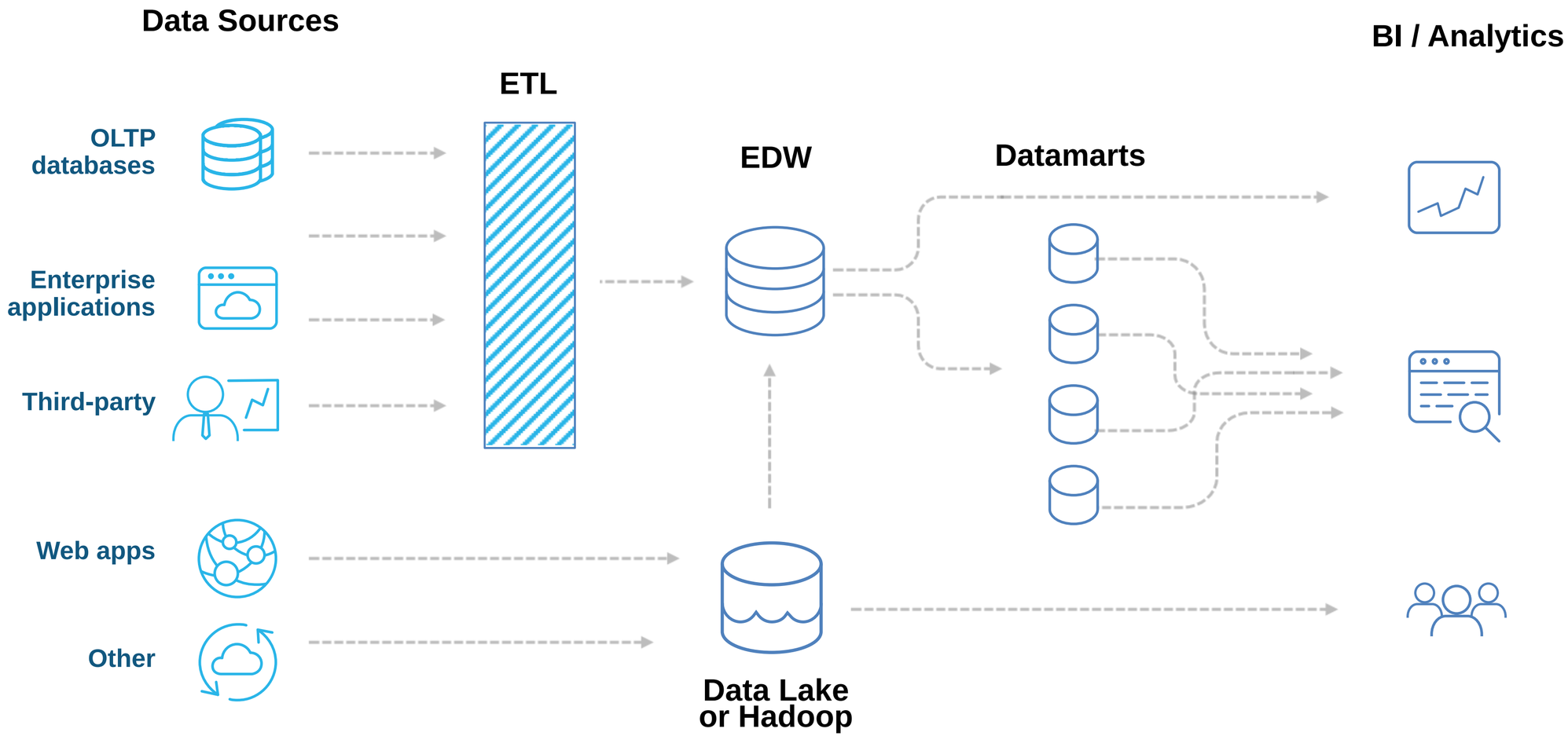
insight
know your numbers

Use Cases

First customers in the region in 2018

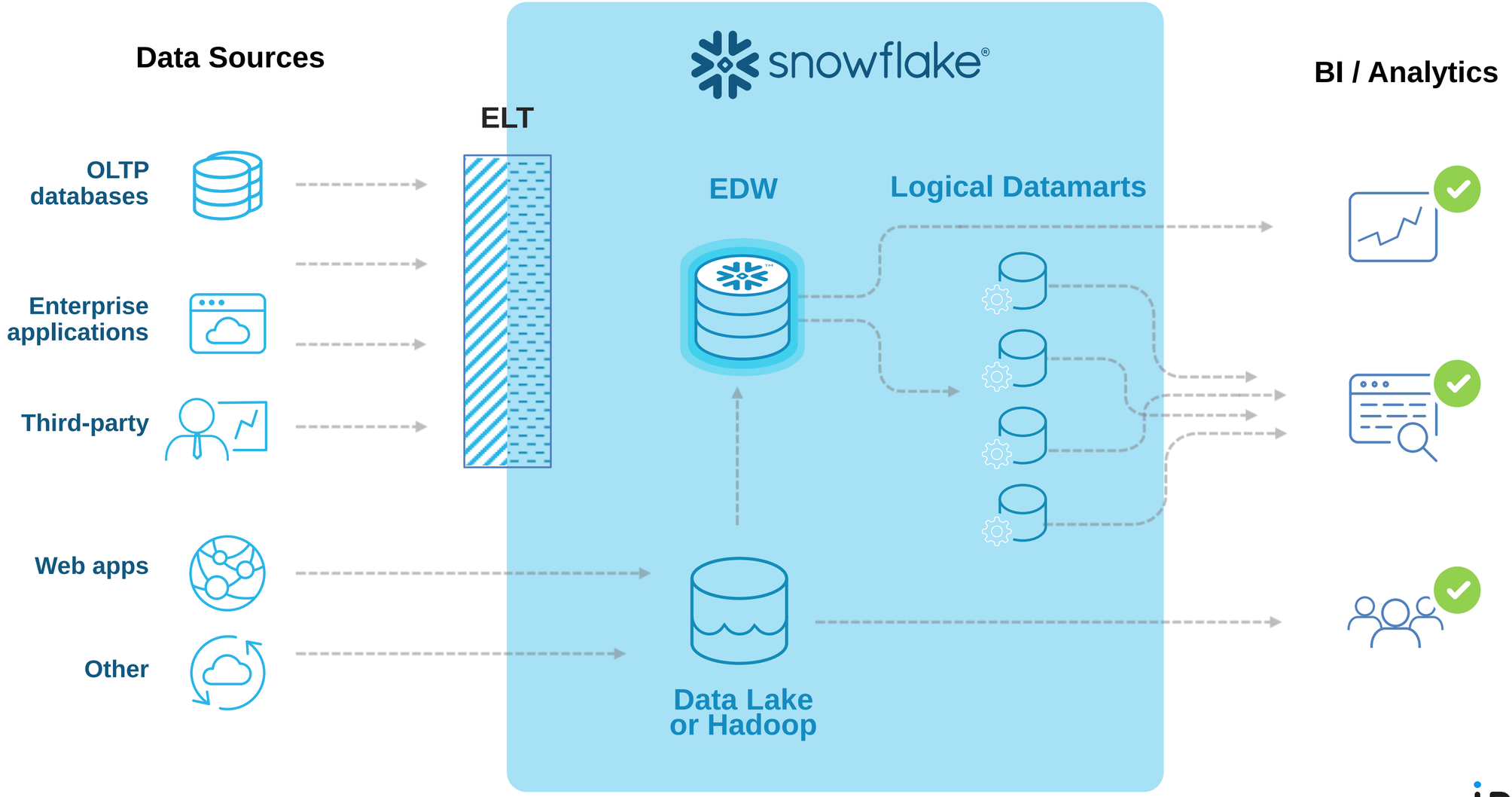
- Petrol Group as a large regional cross industry client with “all-in”
  - In-house development (IBM Mainframe, SAP, Salesforce, custom applications for energy trading on MS SQL, 3rd party sources)
- Government owned insurance company
- Hospitality client with multiple data sources
- Cases for cloud applications integration with an intermediate cloud data warehouse (one day of work and then transfer to on-premise existing data warehouse)
- Multiple POCs for different start-ups and companies

Traditional data landscape



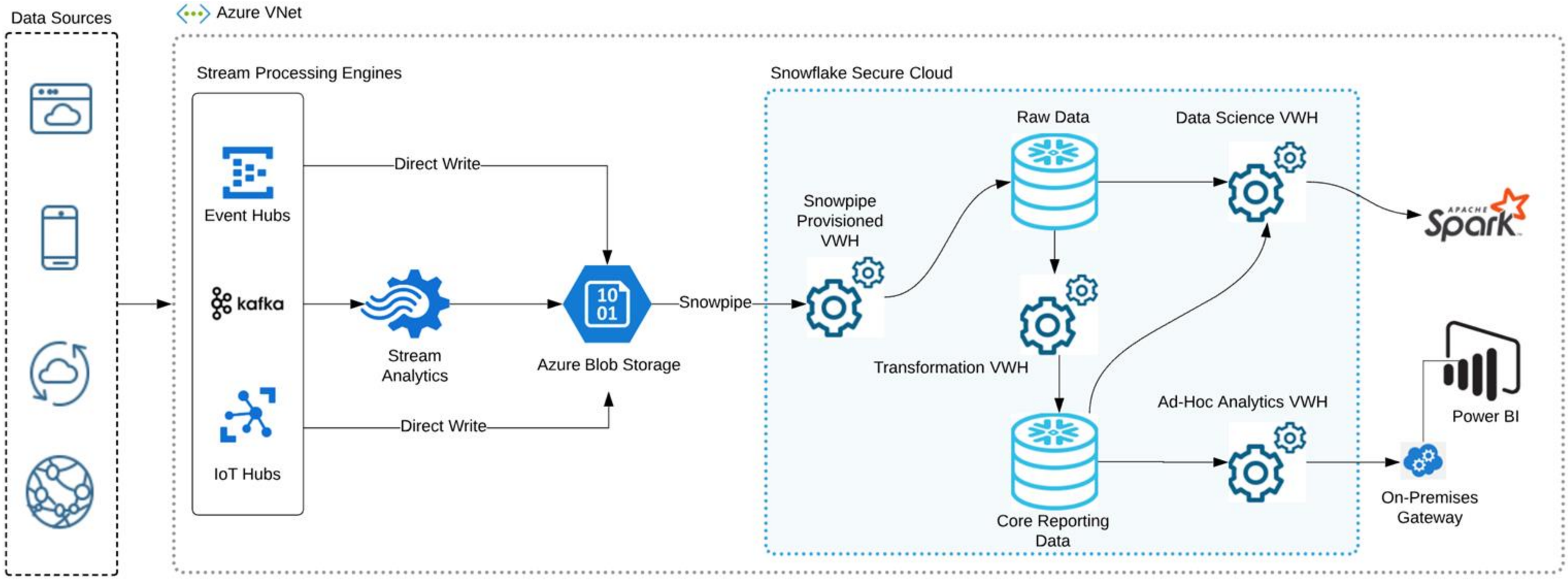
insight
know your numbers

Data Warehouse Modernisation



insight
know your numbers

Streaming Data Analytics and IOT cases



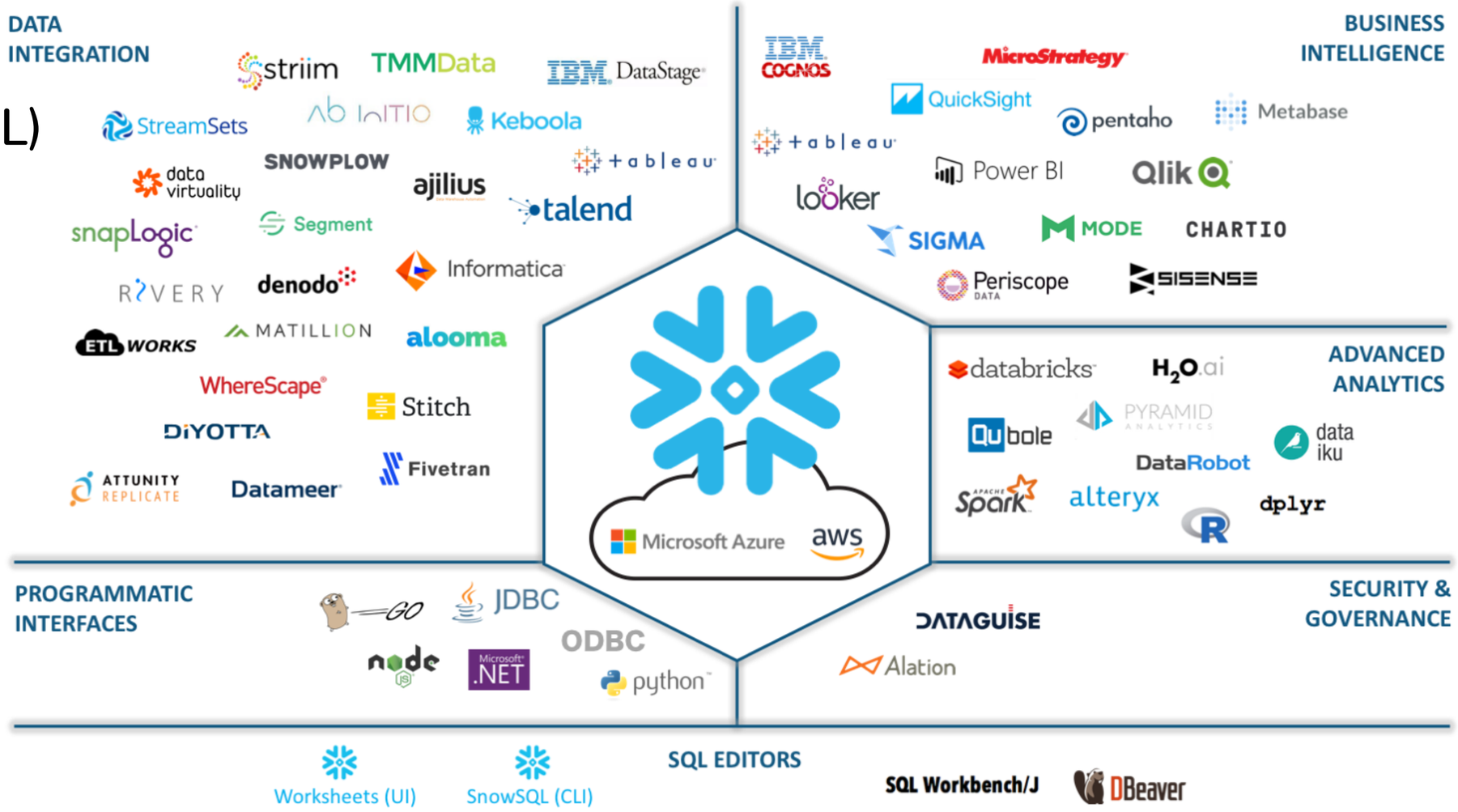
insight
know your numbers

Community

CLI client (SnowSQL)

Connections

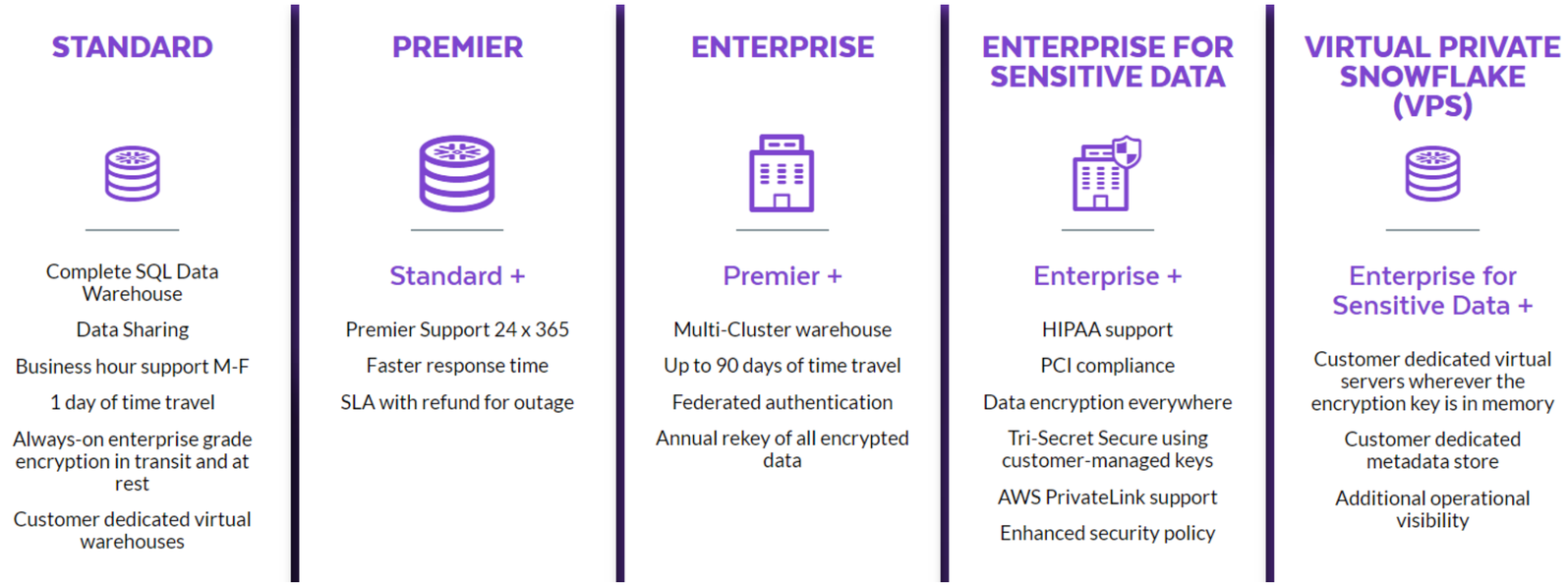
- JDBC
- ODBC
- Phyton
- GO
- Node.js
- .NET
- Spark
- H2O
- DataRobot



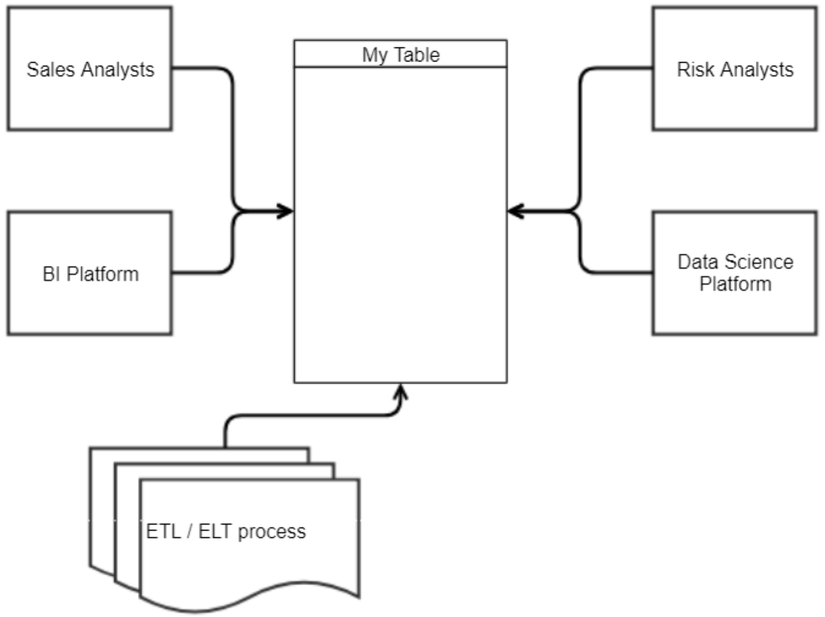
insight
know your numbers

Security

- Fine-grained, role-based access control for data and actions.
- Always-on encryption of data stored in Snowflake.
- Automatic data protection against accidental or intentional destruction.
- Azure ExpressRoute connectivity to utilize Snowflake without going over the public internet.



Changing my axioms



- There are assumptions regarding databases that stood for a long time:
  - Sharing resources
  - Scalability
  - Locking

- The biggest change for me is that there is no more concurrency issue (“unlimited scalability”) and that you can do some thing that ware consider impossible now quite simple☺
  - All your users on one data warehouse with no performance penalty
  - Clone your production environment in seconds
  - Almost no administration
  - Create multi parallel loads against the same table

Snowflake Test Drive

SNOWFLAKE TEST DRIVE

Still having doubts? Arrange a risk-free test drive and try running queries against your own data inside Snowflake to measure performance improvement.

WHAT IS INCLUDED?

- 400 USD credit for Snowflake
- 3 man-days of our services

GET STARTED FOR FREE

HOW DOES IT WORK?

IN516HT — YOU

1. Discovery workshop: 2- to 4-hour session to gain an understanding of your organisation's goals and challenges, get familiar with technologies you use and help you identify the key value opportunity for proof of concept.
2. Preparation of the relevant data set.
3. Preparation of the Snowflake environment.
4. Data Load from your source to Snowflake.
5. 2-hours of Snowflake training.
6. Querying the data and enjoying high performance.

insight know your numbers

ZADENITE BREZPLAČNO LETALSKO VOZOVNICO

LJUBLJANA ✈ BERLIN

IN516HT
know your numbers

Za več informacij obiščite naš razstavni prostor.

IN516HT
know your numbers