Cervello
an ATKearney company

Slim Baltagi
Director, Big Data & ML

DFW Area Advanced Analytics Meetup

Modern Data Warehouses In The Cloud

Use Cases + Live Demo

19th February 2019

Agenda

1. Quick Intro: Talk Format, Speaker Background, Q&A session
2. Key Terms: Cloud, Data Warehouse, Modern, …
3. Data Era: Challenges & Opportunities
4. Traditional Data Warehouses: Key Challenges
5. Modern Data Warehouses: Examples, Solutions, Key Opportunities,
6. Use Cases + Live Demo
7. Key Takeaways & Where To Go From Here?
8. Interactive Session: Q&A, discussion and more networking

1. Quick Intro: Speaker, Talk, Q&A session

- Format of this talk: 40 minutes Talk, a 15 minutes Live Demo, a 30 minutes Q&A session
- Currently
  - Advocating Modern Data Architecture (MDA) and Modern Data Warehouses in the cloud to organizations as director at Cervello Inc., an A.T. Kearney fast-growing professional services company
  - Working with our team of creative problem solvers to deliver projects that help our clients win with data
  - Enjoying emerging technologies (Kubernetes, Kubeflow …) , speaking at conferences and organizing meetups
- In the past
  - Created a mathematical formula for provisioning Apache Hadoop, now an obsolete technology!
  - Evangelized Apache Spark and was the first to fully explain how it relates to Apache Hadoop without any marketing fluff!
  - Initiated a discussion at Hortonworks (recently merged with Cloudera) on its official stance on Apache Spark at a time when Hortonworks itself was happier with Tez and touting it as the future compute engine of Hadoop!
  - Publicly shared the limitations of Spark Streaming. As an alternative, evangelized Apache Flink and introduced it to the western hemisphere. In particular, at Capital One as first adopter before the likes of Uber, Netflix, Lyft, etc … Gave Apache Flink talks in 4 continents
  - Evangelized Apache Kafka as a platform for building end-to-end streaming data applications using Kafka Core, Kafka connect and Kafka Streams/KSQL
  - Delivered many end-to-end data projects as director, architect and engineer

2. Key Terms: Cloud

- What is exactly the cloud? According to Gartner, ‘A style of computing in which scalable and elastic IT-enabled capabilities are delivered as a service using Internet technologies’
- What are the models within the cloud and how do they apply to data warehousing?

	Hardware: Datacenter	Software: Data warehouse	Management: Optimizing, tuning, maintenance	Cook, Eat, Clean
On-premises	You	You	You	At home
Infrastructure IaaS Example: EC2	Vendor	You	You	At a vacation home
Platform PaaS Example: Redshift, EMR	Vendor	Vendor	You	At a fast food restaurant
Software SaaS Example: Snowflake	Vendor	Vendor	Vendor	At a full service restaurant

2. Key Terms: Data Warehouse, Modern

- What is a data warehouse? 'A subject-oriented, integrated, time-variant, non-volatile collection of data in support of management's decision-making process.' Source: Building the data warehouse, a book by Bill Inmon, who is considered to be the father of data warehousing
- From the beginning data warehousing was about the business making better and quicker decisions
- A key factor driving the evolution of data warehousing is the cloud. An updated definition of a data warehouse in this modern era of cloud, social networks and mobile would add attributes such as elastic, scalable (All data, All users),shareable, agile, conductive to exploratory analytics. What do you think?
- Be aware that a traditional data warehouse hosted in the cloud is not necessarily a data warehouse 'built for the cloud' but it is a 'cloud-washed' one! Amazon Redshift would be a good example. Hold on, more to come on this!
- Often modernization in the context of data warehouse is confined to just a 'technology refresh'. How about getting any improvement to the underlying business processes? That would require a 'mindset refresh'!

3. Data Era: Challenges

Regardless of using databases, data lakes, data warehouses or data swamps, businesses all share common challenges related to data in this new era!

- Data is a game changing asset but most organizations struggle to derive business value from it
- Business decisions are hindered by slow and incomplete data analytics
- Legacy technologies, skills, methods and mindsets are stalling innovation
- Changing business requirements are outpacing the capability of IT teams to deliver
- Silos of diverse data from diverse sources (internal enterprise apps, public agencies, external events, 3rd party services, web, …) are more and more segregated, data can not be combined together and business value can not be derived from it
- Costly and complex data infrastructure consuming significant resources to build and maintain data platforms
- Technology selection is becoming more and more challenging due to the abundance of data tools, platforms and services

3. Data Era: Opportunities

What are some of today's opportunities in relation to data in this new era?

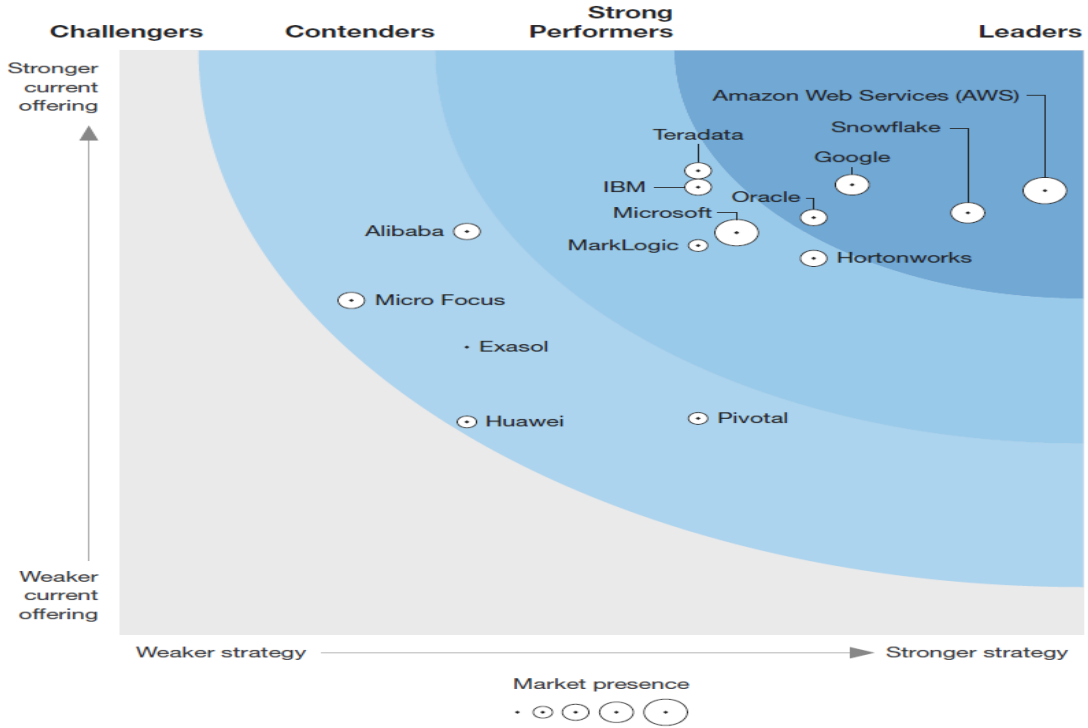
- Cloud computing and the increase in computer-processing power, storage and data-gathering abilities enable businesses to rent much-higher-capacity computer infrastructure from third-party vendors, avoiding the hefty costs and additional resources needed to run their own data centers
- External data sources such as social media data, weather data, demographic data and other third-party data can enrich and augment internal data to help businesses with better decision making
- Real-Time or Near-Real-Time insights that match fast paced business and markets thanks to the maturity of stream processing technologies
- Data-driven decisions thanks to growing adoption of Machine Learning and AI technologies and services and also availability of more powerful compute and cheaper storage services
- Being a data-driven company is not confined to tech giants anymore

4. Traditional Data Warehouses: Key Challenges

1. Scalability: growth in data volume, number of users and applications
2. Concurrency: as number of users increase, they can not operate simultaneously
3. Performance: slow running queries, …
4. Resilience: data backup/retention and node failure protection
5. Complexity: from initial implementation to ongoing maintenance
6. Lack of native support for semi-structured data: JSON and XML formats are popular for data exchange. In traditional data warehouses, data needs to be transformed first and schema needs to be defined before loading
7. High maintenance overhead in the form of constant indexing, tuning, sorting
8. Handling workload fluctuation: sizing servers for workload peaks and valleys
9. Waste of resources: expensive computing resources are wasted during off peak usage
10. Lack of support for processing streaming data
11. Upfront costs: hardware, software licenses, staffing, …
12. Project delays due to provisioning infrastructure

Sure, you are familiar with real world scenarios such as End of Month Reporting, Intensive Load Process, Demanding Executive Dashboard Users, …

5. Modern Data Warehouses: Examples



Reference: *The Forrester Wave™: Cloud Data Warehouse, Q4 2018,* October 29, 2018
The 14 Providers That Matter Most And How They Stack Up

5. Modern Data Warehouses: Examples

Cloud Data Warehouse Platforms

5. Modern Data Warehouses: Solutions

- Major modern data warehouses are cloud based. Most of them do overcome the key challenges of the traditional data warehouses. Unfortunately, overcoming these challenges is happening at different degrees and it is not always a simple check of a yes or no!!

1. Scalability: Scale up, down, or off quickly without delay. Yes for Snowflake. Low for Redshift, Yes for Azure Data Warehouse, Yes for Big Query (but with limits)
2. Concurrency: Lots of users can operate simultaneously
3. Performance: processing bottlenecks and delays, slow running queries, …
4. Resilience: data backup/retention and node failure protection
5. Complexity: Cloud data warehouses are easier to use than on-premise data warehouses
6. Lack of native support of semi-structured data: Although cloud data warehouse do support semi-structured data such as JSON. This support varies from one data warehouse to another.
   - Concurrent throughput for JSON: Yes for Snowflake, No for Redshift, No for Azure Data Warehouse, Moderate for Big Query.
   - High performance for JSON scans: Yes for Snowflake, No for Redshift unless you use the additional Amazon Spectrum tool, No for Azure Data Warehouse, Moderate for Big Query.

5. Modern Data Warehouses: Solutions

7. High maintenance overhead is not a major issue with managed infrastructure for cloud data warehouses, but the level of maintenance varies from one data warehouse to another. This fluctuates to creating indices and distribution keys to near zero management

8. Handling workload fluctuation: With elasticity, cloud data warehouses adapt to workload peaks and valleys

9. Waste of resources: No more wasted resources during off peak usage! Auto suspend, Auto resume? Anybody knows about these features from Snowflake data warehouse?

10. Lack of support for streaming data: For example, loading and processing is possible with Snowpipe, a serverless compute service from Snowflake data warehouse

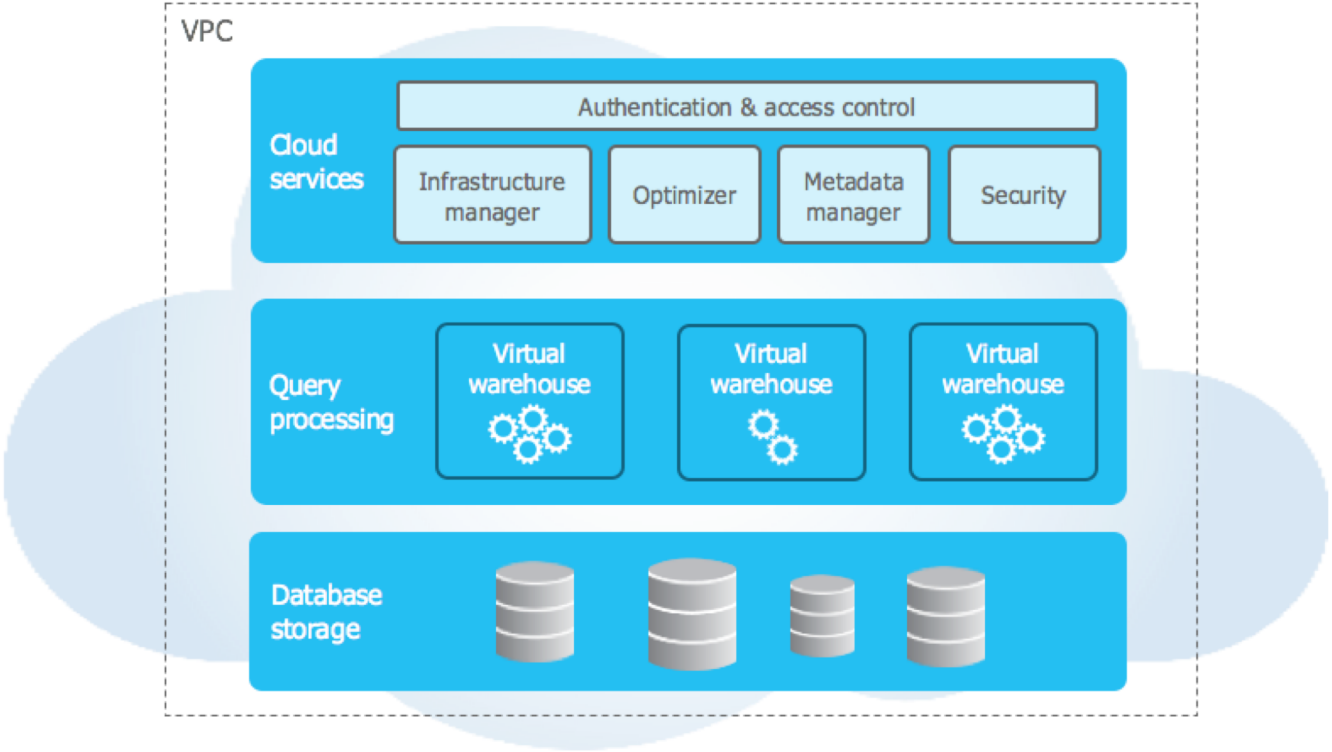
11. Upfront costs: No upfront costs as the model of cloud data warehouse is pay as you go

12. Project Delays due to provisioning infrastructure: Not Applicable for cloud data warehouse. With instant provisioning, projects don't need to wait for infrastructure

5. Modern Data Warehouses: Snowflake key innovations

- Snowflake is based on three key innovations:
  1. Unique architecture : a unique architecture designed for the cloud and able to provide complete elasticity for all your concurrent users and applications
  2. Database engine that natively handles all your data both semi structured and structured data without sacrificing performance nor flexibility
  3. Technology that eliminates the need for manual data warehouse management and tuning. No indexing, tuning, partitioning or vacuuming after loading data => effortless management
- Snowflake architecture consists of three layers, each one is physically decoupled from the other layer and scales independently:
  1. Data Storage layer uses cloud storage to store all data loaded into snowflake in a scalable and inexpensive way
  2. Compute layer comprises of virtual warehouses compute resources that execute data processing tasks required for queries. The virtual warehouses have access to all of the data in the storage layer
  3. Cloud Services layer coordinates the entire system managing security, optimization and metadata

5. Modern Data Warehouses: Snowflake's multi-cluster, shared data architecture



6. Uses Cases + Live Demo

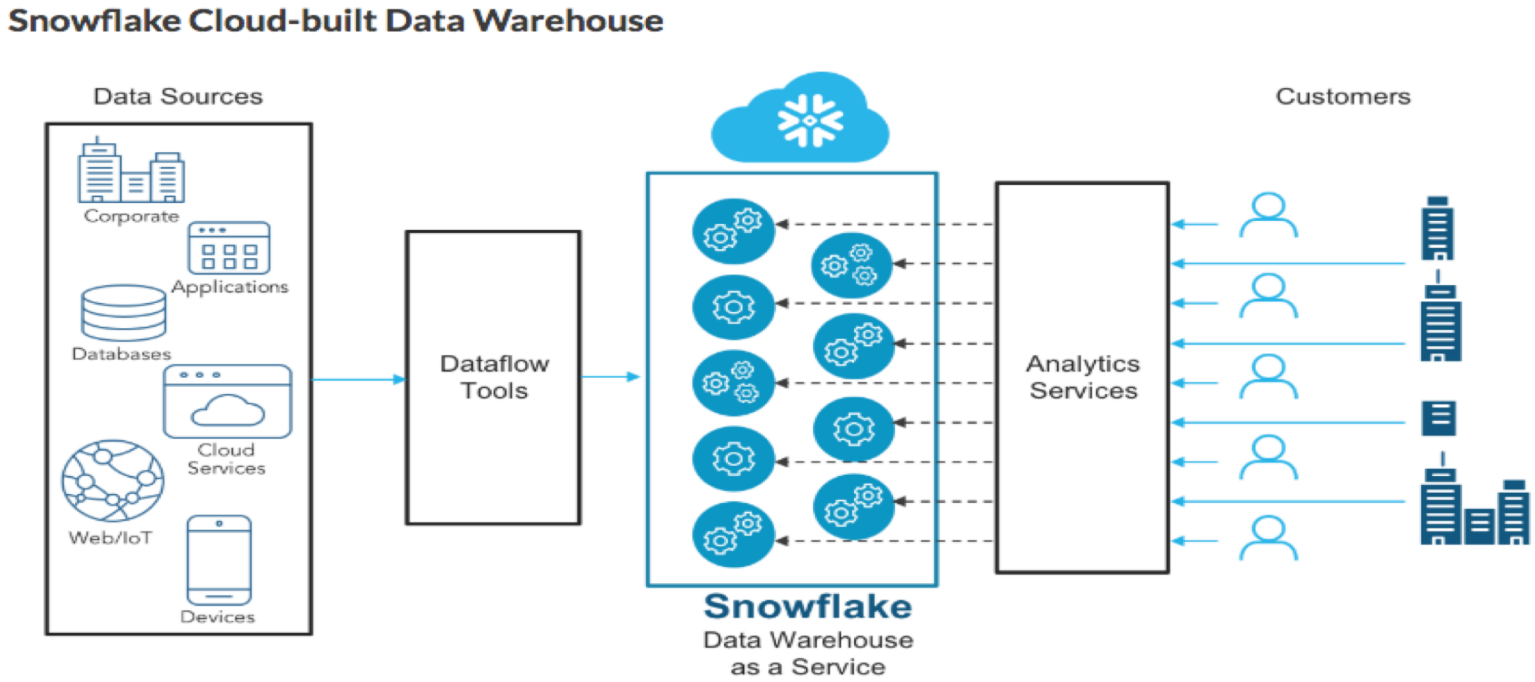
New Use Cases

- Data as a Service
- Self-Service Analytics
- Advanced Analytics Lab
- Real-Time Insight
- Single View of Customer

Data Consumers

- Partners, Suppliers, and Vendors
- Executives, Managers, and Analysts
- Data Scientists
- Embedded Applications
- Many Business Users

6. Use Cases + Live Demo: Snowflake reference architecture



Data Sources/Data Providers: On-premise or Cloud, Structured or Unstructured Data
Dataflow Tools: Streaming Data Platforms, ETL and ELT platforms. E: Extract, L: Load, T: Transform
Data Storage: S3/Azure Storage (missing in the architecture diagram!)
Analytics Services: BI, UI, Data Science, …

6. Use Case + Live Demo: Live Tracking of CTA Buses

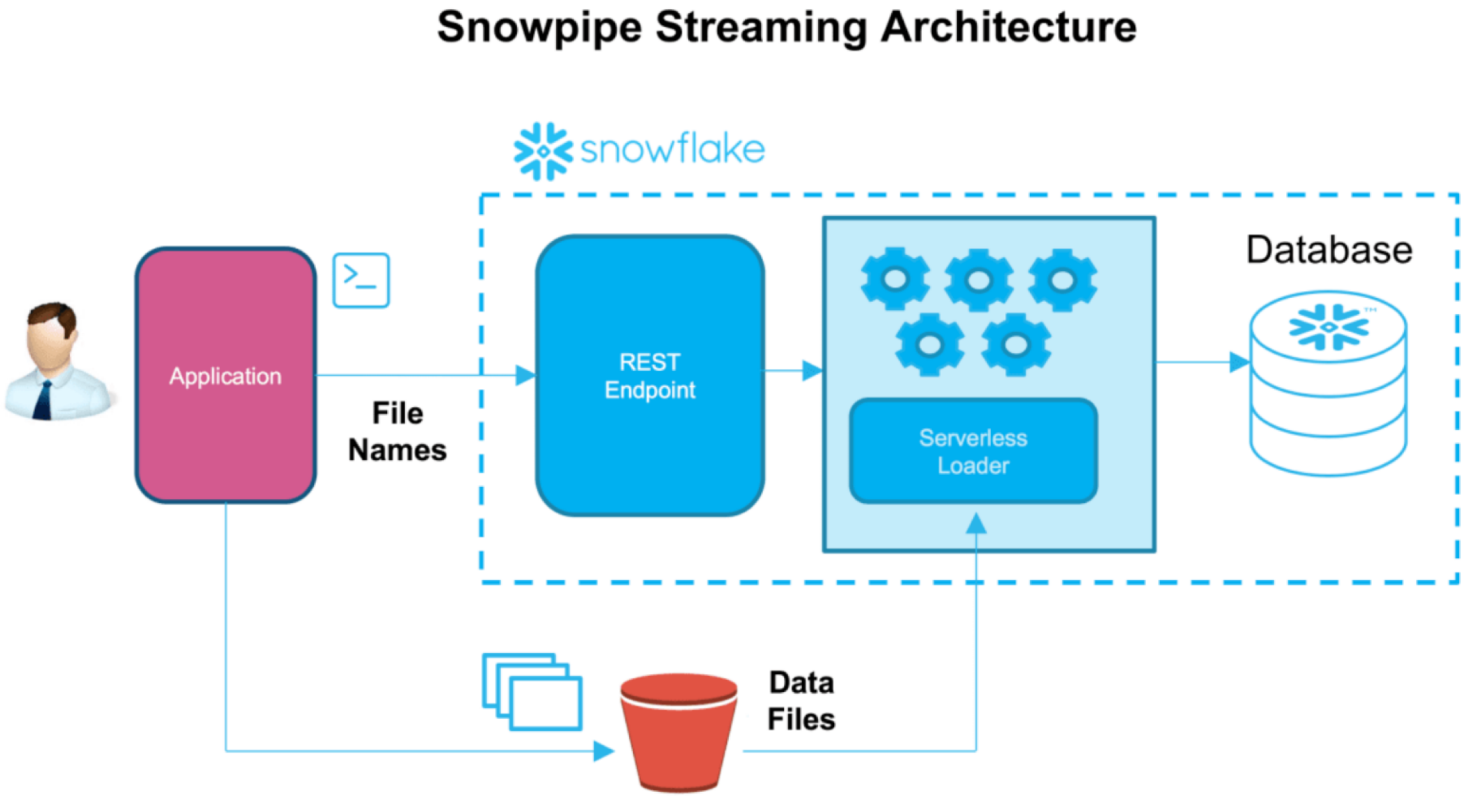
- The purpose of this live demo is to showcase a few aspects of a Snowflake such as Ease of use, Fast provisioning, Native support of semi-structured data, Continuous data loading , …Due to time constraints, we won't cover other unique features of Snowflake such as multi-warehouse concurrency against same data, data sharing, time travel …
- This live demo itself is about an end-to-end application for live tracking of CTA buses that:
  1. Sends a request every minute to CTA Bus Tracker web service about live bus locations and gets an immediate response as JSON data load
  2. Stores the JSON data into Amazon S3 and trigger a notification event into an SQS queue
  3. Consumes the event from the SQS using Snowpipe, a serverless compute service from Snowflake. Snowpipe automatically loads JSON data, as is with no ETL, from S3 into a Snowflake SQL database
  4. Queries the JSON data using Snowflake's flatten method as extension to standard ANSI SQL
  5. Visualizes the buses locations on a map
- This is not a toy use case! Since JSON is a popular format for data exchange, the same data pipeline can be leveraged in many other use cases with different data sources: internal applications, external enterprise data services, machine generated data such as IoT devices …

6. Use Case + Live Demo: Live Tracking of CTA Buses

- What did I use to cook this application?

  - AWS Free tier: a free account from https://aws.amazon.com/free/
  - Snowflake Test Drive: 30 days free trial through our partnership with Snowflake. Who wants one? https://trial.snowflake.com/?utm_source=cervello&utm_medium=referral&utm_campaign=self-service-partner-referral-cervello
  - CTA (Chicago Transit Authority) Bus Tracker API to get free up-to-the-minute JSON data feeds: https://www.transitchicago.com/developers/bustracker/ You just need to request a free API key from CTA
  - Anaconda: Free and open source distribution of Python and R programming language. Comes with Python IDE (Spyder 3.3.2) and other goodies!
  - Free trial of Tableau Online: https://www.tableau.com/products/cloud-bi

6. Use Cases + Live Demo: Live Tracking of CTA buses



7. Key Takeaways & Where To Go From Here?

- Try it yourself
  - Snowflake is worth a Free trial! https://trial.snowflake.com/?utm_source=cervello&utm_medium=referral&utm_campaign=self-service-partner-referral-cervello
  - You can build a short term POC in either **AWS** or **Azure** while using US $400 credit for 30 days for both compute and storage,
- Engage the business
  - It is critical to identify business sponsorship and use cases
  - Cervello prefers short, high impact workshop approach
- Get end to end help
  - Pick the right data platform … For example, the most 'popular' data warehouse is not necessarily the one that fulfills your data warehouse needs! We helped many of our clients move from Amazon Redshift and other traditional data warehouses to Snowflake.
  - Avoid having a pure 'technology refresh' and missing the opportunities to add business value (refer to use cases)
- Get advice

8. Interactive Session: Q&A, discussion and more networking

- The goal here is for the audience to chime in now and later on the comments section of the event or the meetup discussions section.
- Here is a few samples of food for thought!
  - Can somebody briefly share his/her real-world experience migrating from a legacy data warehouse to a modern data warehouse?
  - Did anybody worked on a fit-for-purpose analysis of modern data warehouses?
  - What would be the best practices of using a modern data warehouse such as Snowflake?
  - What advice will you give someone transitioning his/her career to work on a modern data warehouse?
  - Is there anyone who would like to give a follow up talk on modern data warehousing?

Thank You!

Learn more about Cervello at mycervello.com

Boston | New York | Dallas | London

Get in touch with contributors:

Slim Baltagi

sbaltagi@gmail.com

https://www.linkedin.com/in/slimbaltagi/

@SlimBaltagi

Jim Leavitt

jleavitt@mycervello.com

https://www.linkedin.com/in/jimleavitt/

Appendix

Cervello Profile

We are a professional services firm focused on helping organizations win with data. We have a global presence with an ability to service customers across industries. We are organized around three practices that our under pinned by our expertise in modern data architectures.

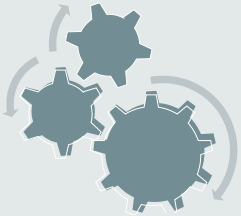
PERFORMANCE MANAGEMENT | **SUPPLIER + CUSTOMER RELATIONSHIPS** | **DATA MONETIZATION + PRODUCTS**

WHAT WE DELIVER

MODERN DATA ARCHITECTURE & PLATFORMS

Using leading cloud technology platforms we develop and design analytics-ready connected data solutions

DATA MANAGEMENT & MACHINE LEARNING



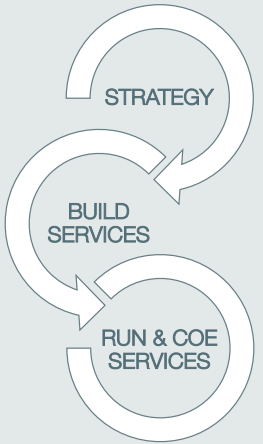
We embed data management processes and use machine learning to improve data quality

BUSINESS DRIVEN INSIGHTS & ACTIONS

- ANALYTICS & DATA SCIENCE
- MODELING & PLANNING
- MOBILE EXPERIENCE

We provide different methods to consume data to facilitate insights and actions inside and outside the organization

HOW WE DELIVER



Our teams are located in Boston, New York, Dallas, London and Bangalore

Cervello win with data. + snowflake

9 Years Working In The Cloud

30 Snowflake Trained Consultants

100+ Big Data & Analytics Engagements

End to End Enablement

tableau · Power BI · looker · amazon web services · Microsoft Azure

Cervello Value Accelerator

PLAN
< 1 WEEK

+ Use case inventory
+ Readiness Assessment
+ Architecture Design

TRAIN
2 DAYS

+ Getting around Snowflake
+ Analyzing & Reporting Data
+ Data Management & Security

ACTION
1-2 WEEKS

+ Use case & MVP defined
+ POC Implementation
+ Stakeholder Presentation
+ TCO Analysis