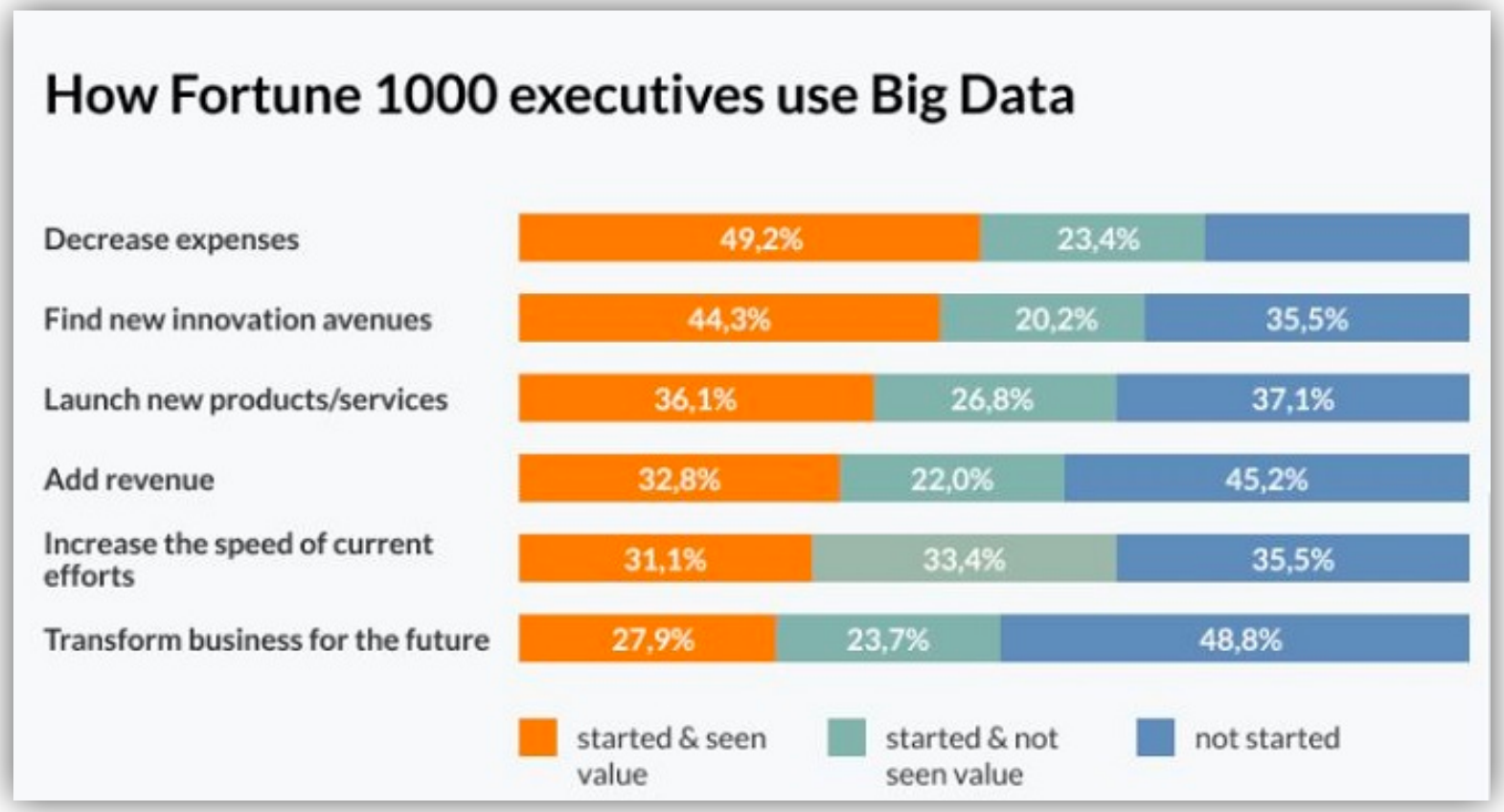
Designing A Big Data Architecture Diagram

Designing A Big Data Architecture Diagram

- How Big Data Analytics Projects Meet Organizational Needs
- Big Data Case Studies
- Best Practices in Analytics Project Design
- Selling Your Project Internally
- Brainstorm A Data-Driven Strategy
- Key Drivers for Analytics Solutions
- Identifying Analytical Opportunities Within Organizations
- Defining and Assessing Problems
- Storage & Compute
- Classroom Architecture Design Presentations

How Big Data Projects Meet Organizational Needs?

Organizational Objectives



Source: [Newvantage Partners Big Data Executive Survey 2017]

Project Charter

Identify The Challenges

- What business challenges need to be addressed?
- What are the drivers of the initiative?

Technical Details

- Data sources needed and available?
- Cloud vs On-Prem?
- Project duration?
- Technologies proposed?

Talent & Resources

- Do you have the talent and resources to allocate to this project?

Key Stakeholders

- Who are the key stakeholders to approve and support this initiative?

Leverage Points

- Are there any other current data initiatives that can be leveraged for this project?
- Any current jobs or analyses that support this project?

Organizations Leveraging Analytics

Scalable Architecture

Is the current architecture designed to scale for future data growth?

Analytics Team

Is the current architecture designed to scale for future data growth?

Are the data initiatives and projects in production running efficient and optimal? Technical debt?

Resource Efficiency

How many data project are currently in production? Production jobs indicate the confidence and health of the project.

Production Jobs

Big Data Case Studies

Personalization analytics

Content delivery **optimization**

Movie tagging

Predict what customers will enjoy watching

Subscriber **profiling**

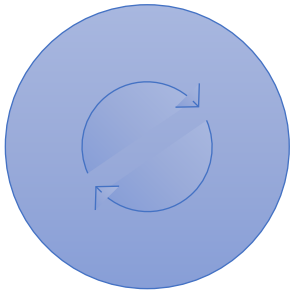
Messaging analytics

- Data collected by users **as they browse** Facebook is used to match them with companies paying for marketing.
- Collect data to use for targeted ads at their own customers.
- Provide services for virtual reality interaction.

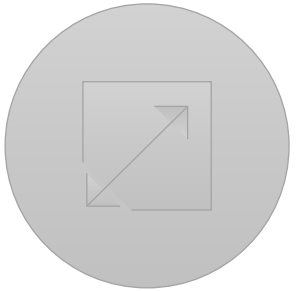
LinkedIn

- Collect **every click**, **page view**, and **interactions**.
- Aid decision making and suggestions for their users based on number user interaction with the site. Suggesting other contacts, other resources and other interests. Pulling Twitter information as well.
- Personalized recommendation based on user behavior and habits: *e.g. Company B vs Company A priority*.
- Stream-processing technology to ensure the most up-to-date information displayed on site.

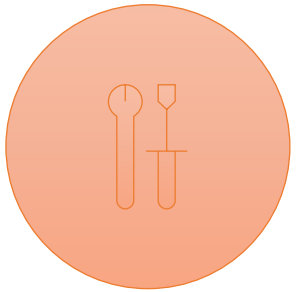
Use Case Breakdown



Identify the challenges and background



Technical Details & Data Sources



Begin designing the architecture diagram

Background

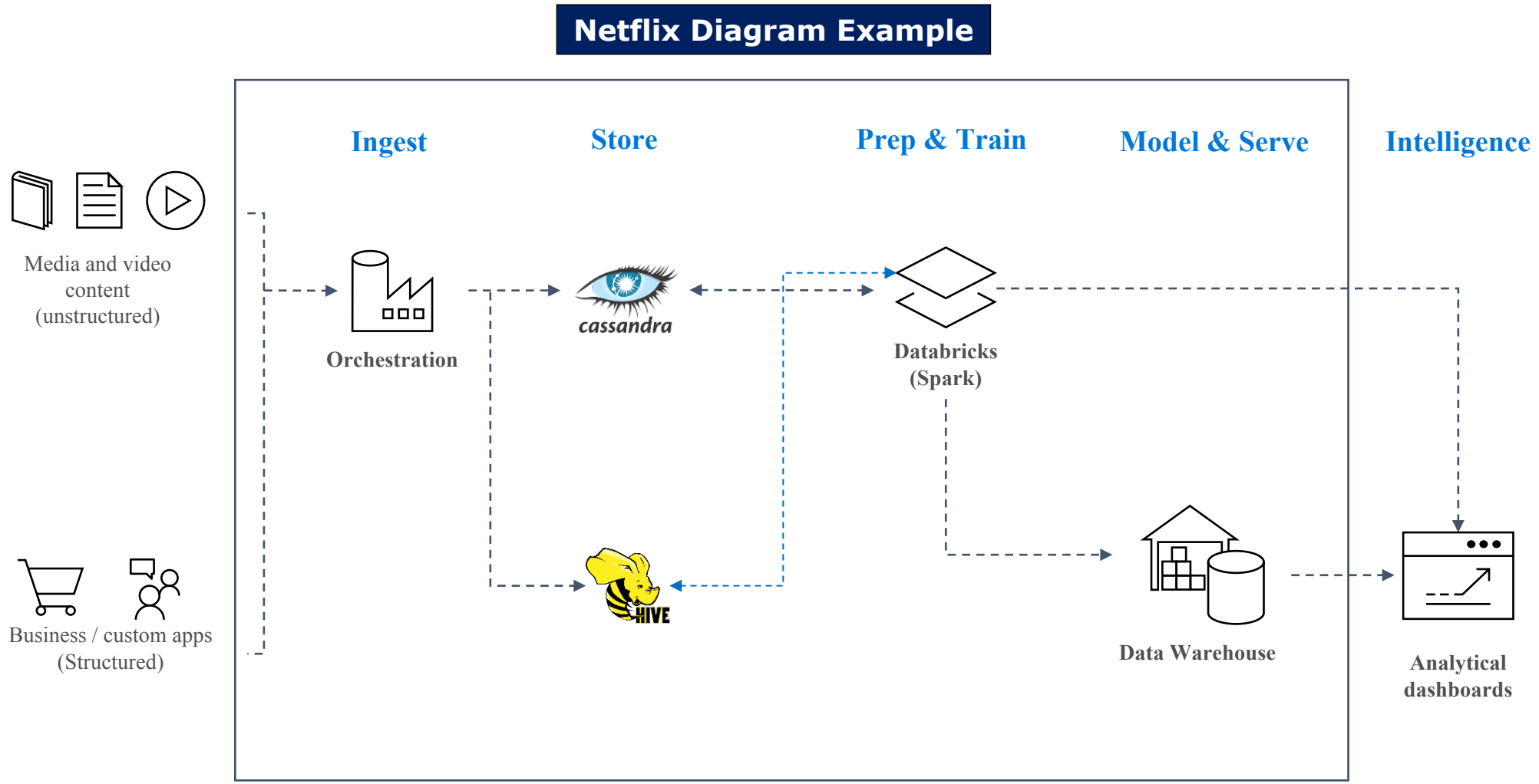
Netflix is a streaming movie and TV service, known to account for a significant amount of internet traffic in the US. The service has over 65 million members in over 50 countries enjoying more than 100 million hours of TV shows and movies a day. Netflix has multiple data initiatives across the organization. Most notably, they focus on personalization analytics, messaging analytics, content delivery analytics, device analytics and more. The holy grail has always been to predict what customers will enjoy watching. Netflix leverages a custom recommendation engine.

Data:

- Titles watched
- Viewing time of day
- Frequencies of playback and stoppages
- Rating given
- Quality of experience
- Rebuffer rate and bitrate

Technology:

- NoSQL
- Cassandra
- AWS
- Spark
- Hadoop
- Hive


Netflix Diagram Example
Ingest
Store
Prep & Train
Model & Serve
Intelligence
Media and video content (unstructured)
Business / custom apps (Structured)
Orchestration
cassandra
HIVE
Databricks (Spark)
Data Warehouse
Analytical dashboards

Background

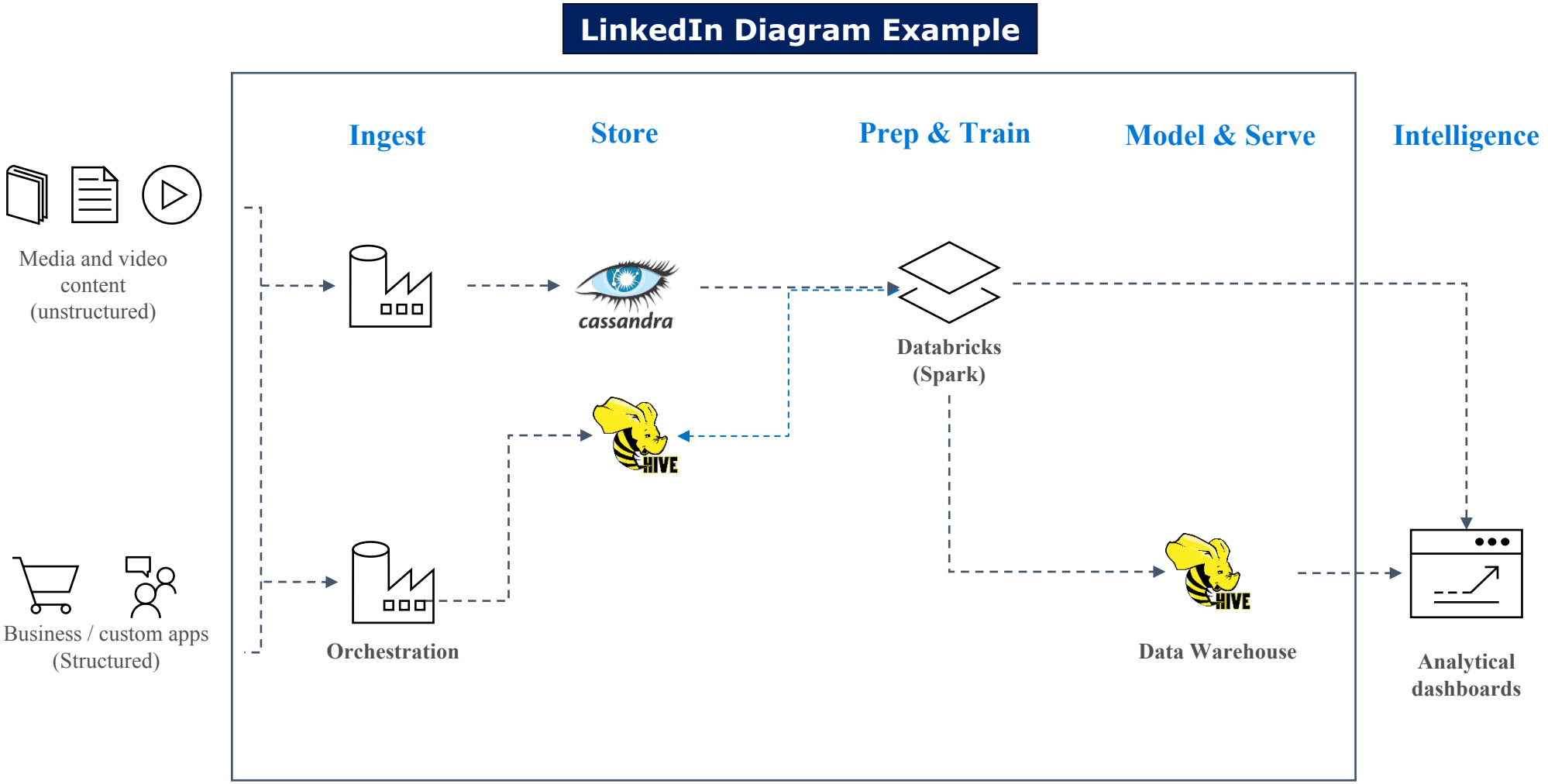
Competition among social networks is fiercer than ever and what's hot one year may not be the next. LinkedIn need to ensure their site remains an essential tool for professionals. As such, Big Data is at the heart of LinkedIn's operations and decision making, helping them provide the best possible service for the sit's millions of members. LinkedIn tracks every move users make on the site: every click, every page view, every interaction. With 410 MM members, that want to provide the ultimate customer experience. Recommending "people you may know", better suggestions for users and allow you to see who has viewed your page.

Data:

- Site clicks, shares, likes
- All contact message
- Serves 10Ks of webpages every second every sate

Technology:

- Oracle DB
- Kafka
- Spark
- Hadoop
- Hive
- MySQL


LinkedIn Diagram Example
Ingest
Store
Prep & Train
Model & Serve
Intelligence
Media and video content (unstructured)
Business / custom apps (Structured)
cassandra
HIVE
Databricks (Spark)
Orchestration
HIVE
Data Warehouse
Analytical dashboards

Background

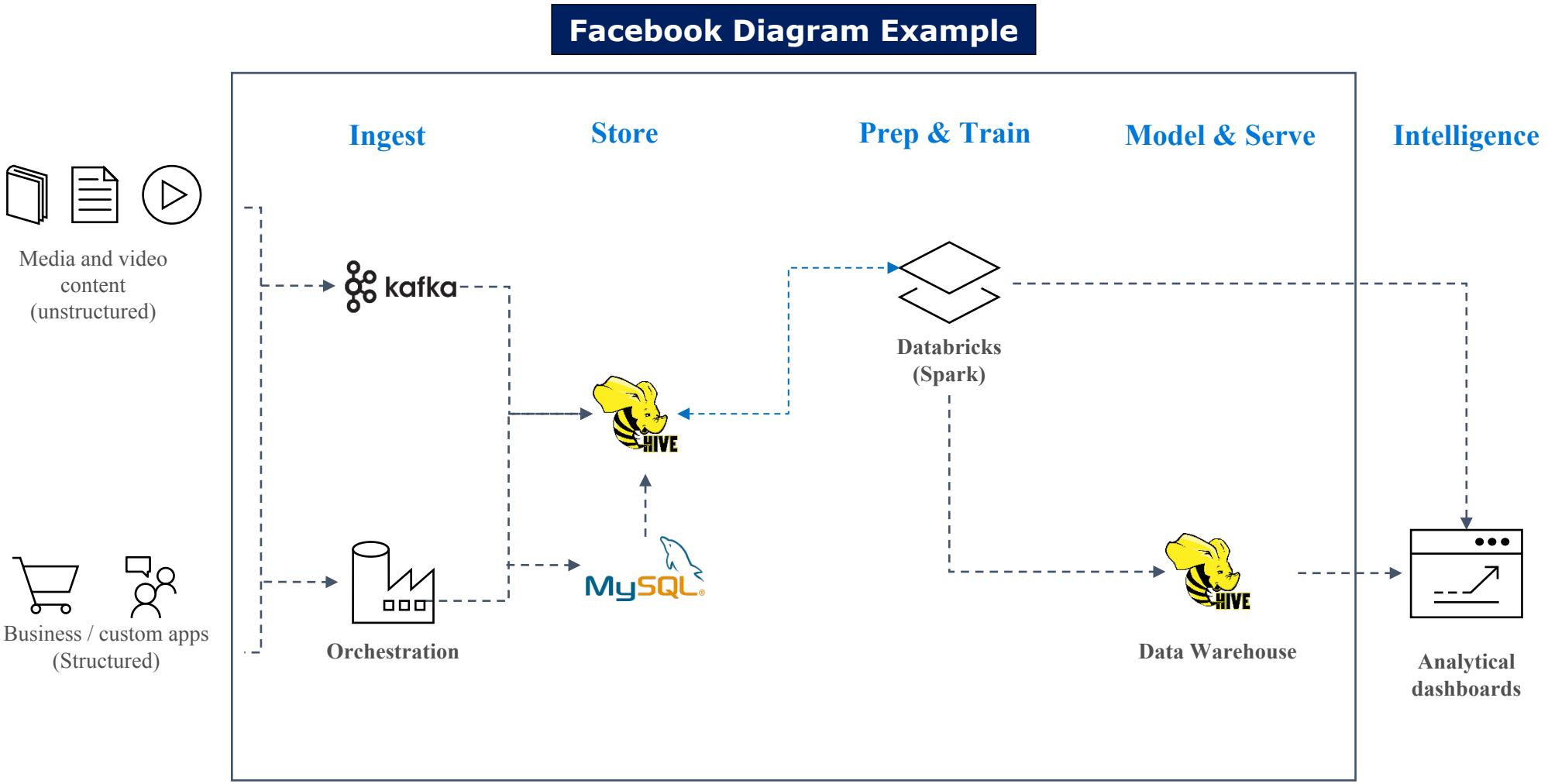
Facebook, by some considerable margin, is still the world's biggest social network. And Facebook, with 1.5 billion active monthly users, has access to far more user data than just about anyone. With access to demographic data, Facebook make a significant amount of money selling ad space. Facebook, together with its users, generates its own data. Facebook leverages the photos uploaded to enhance facial recognition for social selling purposes.

Data:

- Site clicks, shares, likes
- All contact info and demographics
- Uploaded photos and content

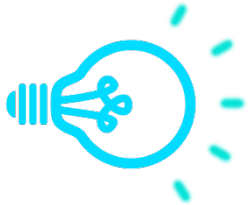
Technology:

- Hadoop
- Hive
- MySQL
- PyTorch
- HBase


Facebook Diagram Example
Ingest
Store
Prep & Train
Model & Serve
Intelligence
Media and video content (unstructured)
Business / custom apps (Structured)
kafka
Orchestration
HIVE
MySQL
Databricks (Spark)
HIVE
Data Warehouse
Analytical dashboards

Best Practices in Analytics Project Design

- ✓ **Insights** to action



- ✓ Drive Revenue

- ✓ Improve **Efficiencies**
- ✓ *Save* Money

- ✓ Start with a Proof of Concept

- ✓ Stakeholder **Visibility**
- ✓ Leverage **Cloud** Partner **Services**

Selling Your Project Internally

Let's reflect on our own organizations.

Do you have the talent and resources to allocate to this project?	5 Minutes	Who are the key stakeholders to approve and support this initiative?
Talent & Resources		**Key Stakeholders**

Case Study: GE

Data collected by **sensors** installed in **machinery** provides information on how it is operating.

Fine-tune the pitch and direction of the blades of their 22K wind turbines to maximize energy captured.

Schedule maintenance and reduce delays or workshop congestion caused by unexpected failures.

Predict when parts are likely to fail and need repairs.

Background

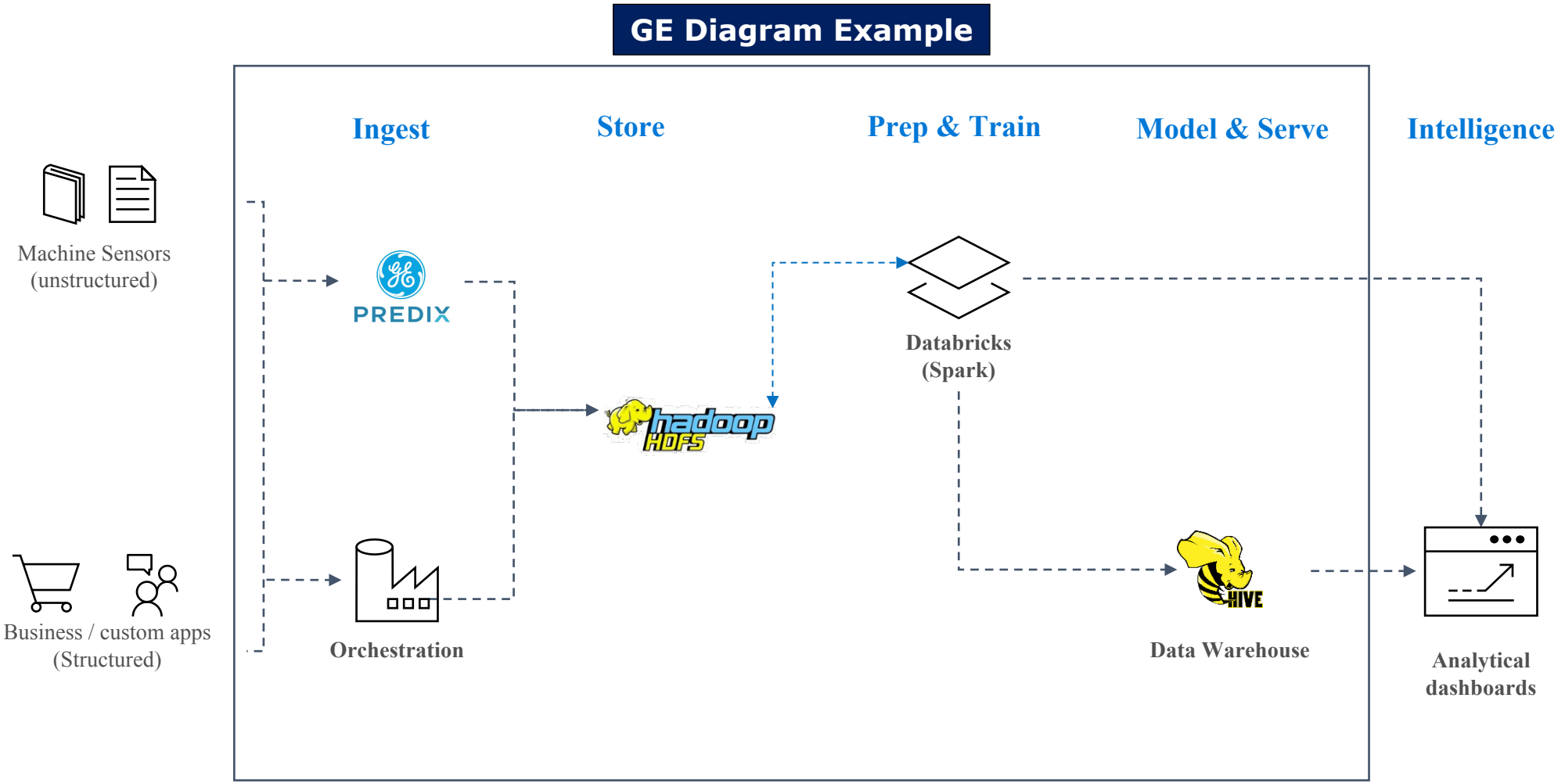
During the Big Data era, GE unveiled plans to create what they call the Industrial Internet. Downtime of essential machinery can directly lead to loss revenue and costly human resources have to be assigned to the upkeep and maintenance of systems. Data collected by sensors installed in machinery across every sector in which GE work is measured and analyzed to provide information on how it is operating. This data is used both internally, but also to sell this sensor monitoring system.

Data:

- Internal machine sensor data
- External data: political, meteorological
- One gas power station turbine generates 500GB/day

Technology:

- Hadoop Data Lake
- Hive
- Predix
- Private Cloud
- Predictivity


GE Diagram Example
Ingest
Store
Prep & Train
Model & Serve
Intelligence
Machine Sensors
(unstructured)
PREDIX
Databricks
(Spark)
hadoop
HDFS
Business / custom apps
(Structured)
Orchestration
HIVE
Data Warehouse
Analytical
dashboards

Key Drivers for Analytics Solutions

Hardware Limitations

Is the current architecture struggling to handle data volume growth?

Limited Analytics Talent

Is the current analytics staff over worked? Can't meet consistent demands of the organization?

Need Revenue Growth

Data driven initiatives provide risk adverse opportunities to grow revenue.

Access To New Information

If your organization recently invested in new operational technology, it is an opportunity to leverage the new data source.

	Cloud	On-Prem
Pros	• Low-cost up front • No hardware/software investments • More consistency and updated	• More control on setup of systems • Direct DB access • Cheaper long term • Security in your hands
Cons	• Spend more long-term • Shared security policy • No direct DB access	• Have to pay for hardware and servers • Challenge for DR • Sole responsible for security

****Hybrid is also an option.**

Identifying Analytical Opportunities Within Organizations

Analytical Opportunities

Available Data

- What current data sources are available today?
- What new technology upgrades have been made recently? Likely a new data source.

Leaky LOB

- Are there any lines of business that are underperforming?
- Any new lines of business?
- Change of management for a given business line?
- New technologies within a business line?

Reflect Competitor

- Observe projects and initiatives of your competitors?

Industry Research

- Research common KPIs for your industry?
- Find similar industry projects that are relatable.

Leverage Points

- Are there any other current data initiatives that can be leveraged for this project?
- Any current jobs or analyses that support this project?

Defining and Assessing Problems

Reflect On The Business Challenges (Session 4)

What business challenges need to be addressed?
What are the drivers of the initiative?

Define Two Business Problems

1

2

10 Minutes

Reflect On The Impact Of The Analysis

If the business challenge is addressed how will that impact your business?

Describe Two Potential Impacts To The Business

1

2

10 Minutes

Available Data Sources

Reflect On The Technical Details

Data sources needed? Data sources available? Cloud vs On-Premise? Project duration? Technologies proposed?

List potential available data sources.

1

Describe the use of the data.

2

10 Minutes

Reflect On Talent & Resources

Reflect On Key Stakeholders

List the source of talent or personnel to support the project.

1

List the key stakeholders to support the project

2

10 Minutes

Reflect On Cloud vs On-prem Options

Choose an environment. Explain why…

10 Minutes

Reflect On The Storage and Compute Options

What are expected data volumes? Growth rate? Where is your environment? How will you ingest data?

List your primary storage options.

1

List your primary compute technologies.

2

How will solution ingest and serve the data?

3

15 Minutes

Org Diagram Example

Sources

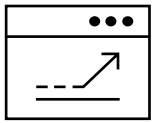
Ingest

Store

Prep & Train

Model & Serve

Intelligence



Analytical dashboards

Participants
Architecture
Design
Presentations

Meirc Training & Consulting

PLUS SPECIALTY TRAINING

Conclusion

Meirc Training & Consulting

PLUS SPECIALTY TRAINING

Big Data is the present and the future…

Are you ready?

"Without data you are just another person with an opinion"

W. Edwards Deming