Introduction to Big Data

ECAP456

Dr. Rajni Bhalla

Associate Professor

Learning Outcomes

After this lecture, you will be able to

- understand what is data stream.
- learn Use Cases for Real-Time and Streaming Data
- understand Data Lake
- differentiate between data lake vs data warehouse

What Does Big Data Streaming Mean

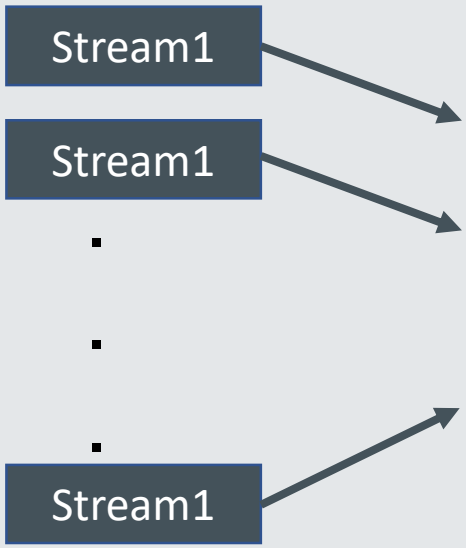
- Big data streaming is a process in which big data is quickly processed in order to extract real-time insights from it.
- The data on which processing is done is the data in motion.
- Big data streaming is ideally a speed-focused approach wherein a continuous stream of data is processed.

What Does Big Data Streaming Mean

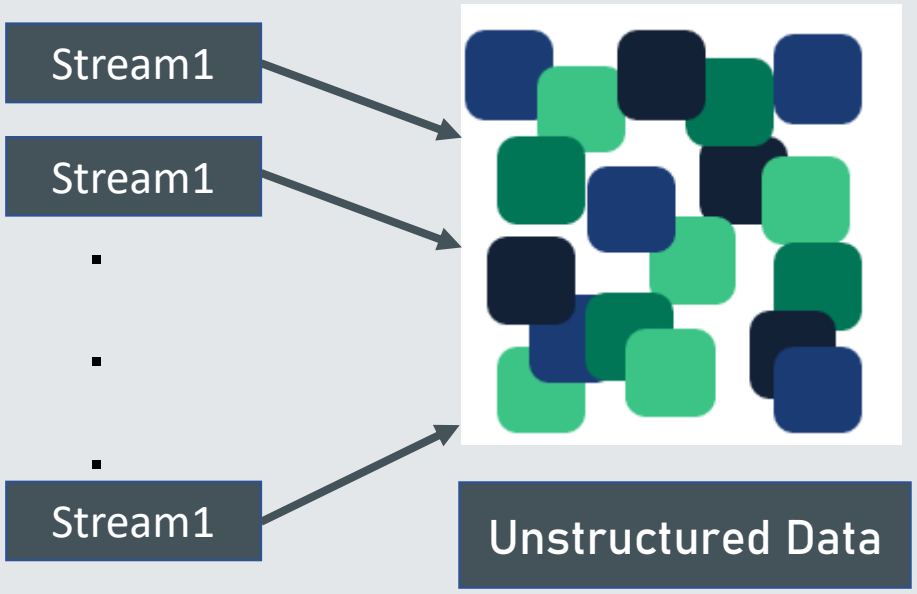
Data Streaming

Big data streaming is a process in which large streams of real-time data are processed with the sole aim of extracting insights and useful trends out of it.

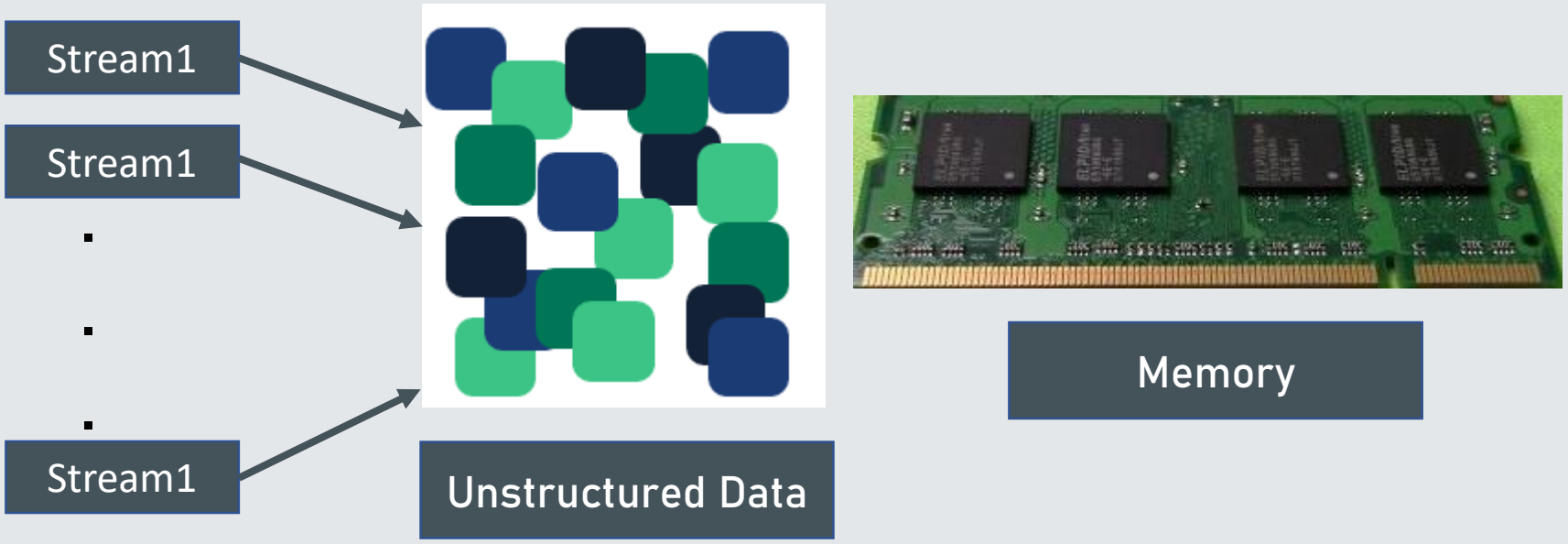
Data Streaming



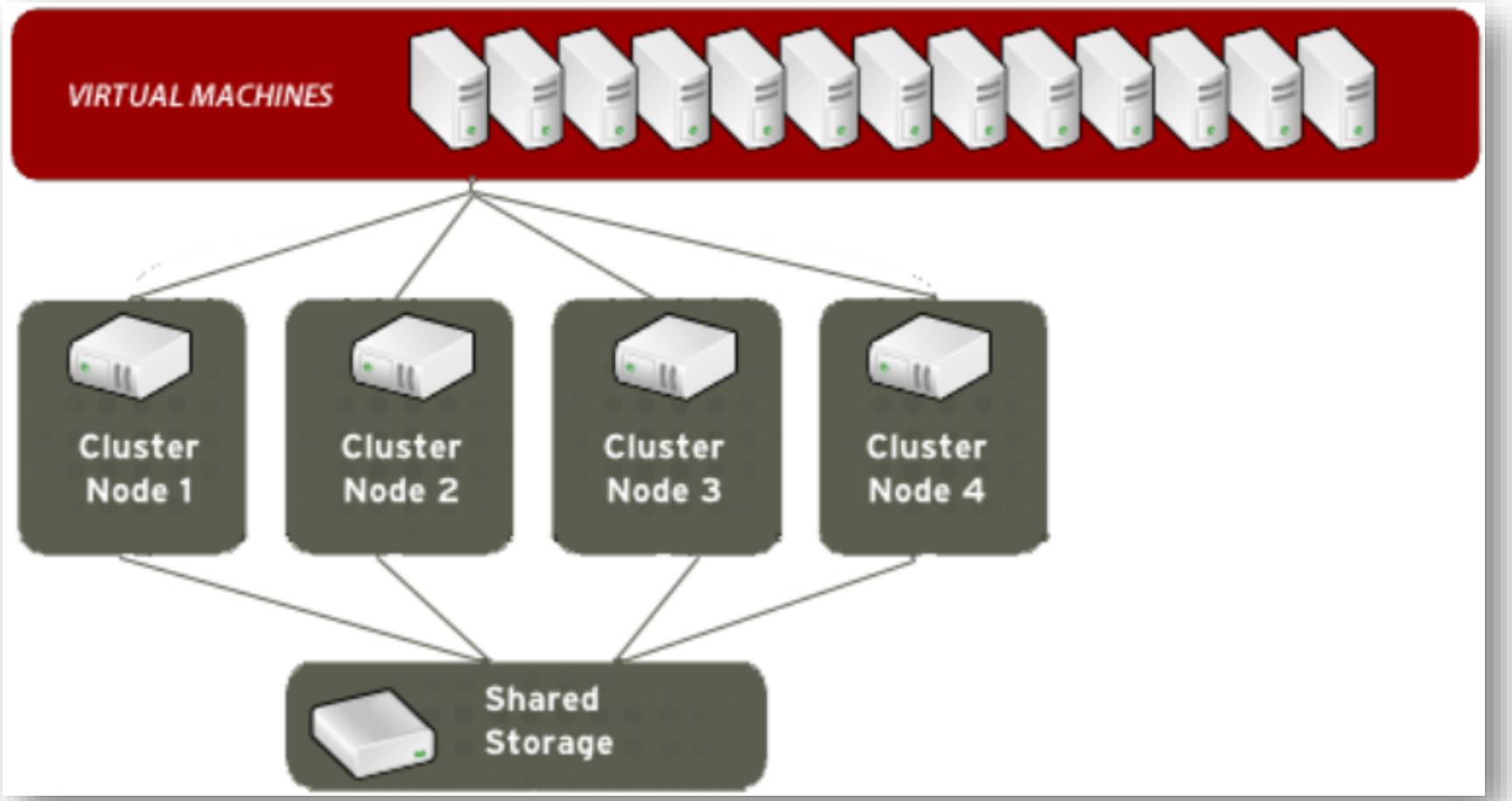
Data Streaming



Data Streaming



Data Streaming



Cluster of servers

Data Streaming

- Speed matters the most in big data streaming.
- The value of data, if not processed quickly, decreases with time.
- Real-time streaming data analysis is a single-pass analysis.
- Analysts cannot choose to reanalyze the data once it is streamed.

Data Streaming

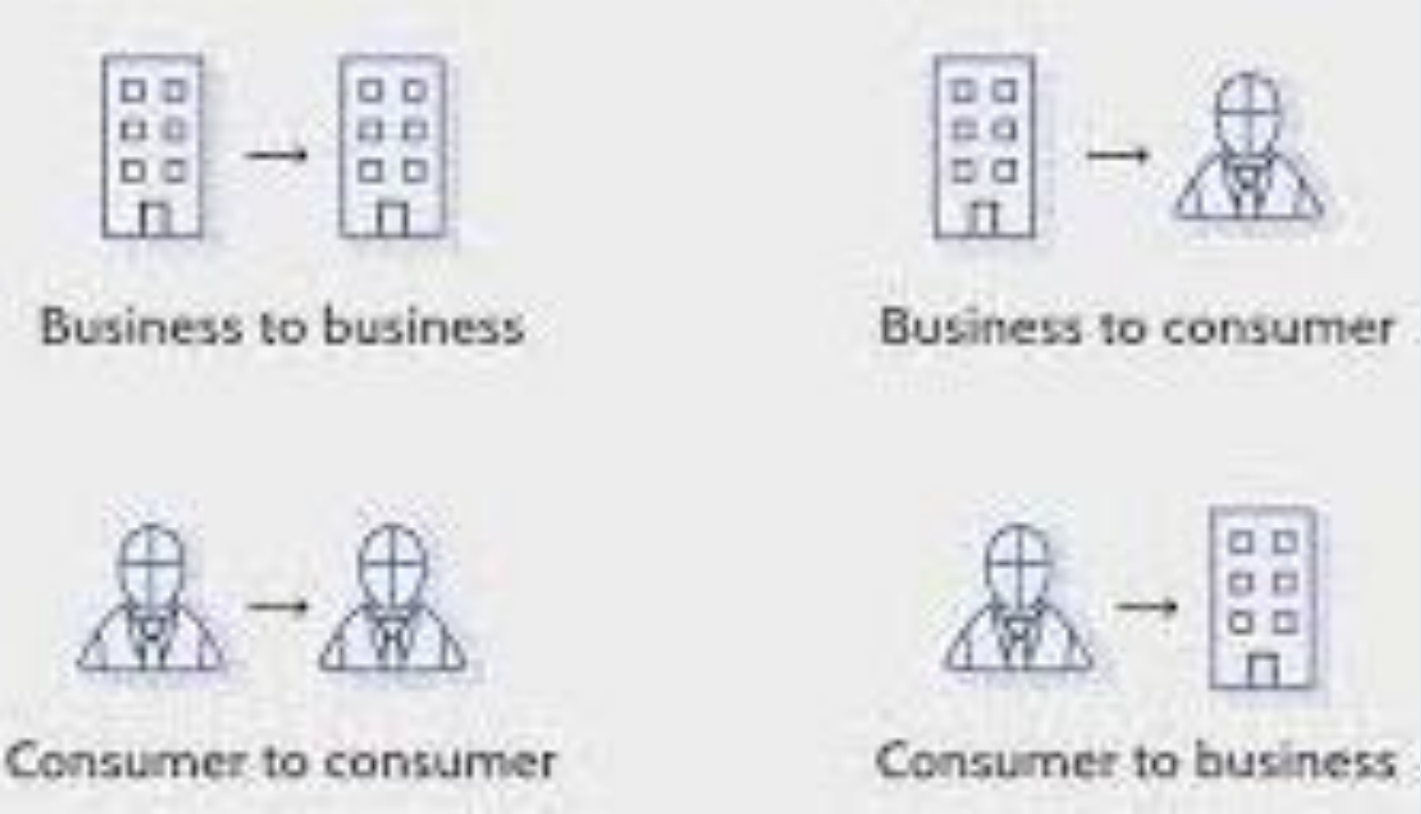
Social Media

Camera or Sensors

Example Where Real Time Streaming Data Is Created (Log Files)

```
<start_timestamp> <end_timestamp> <GMT_offset> <application_name> <machine_IP> <username>
2007-02-13 17:17:28     2007-02-14 12:06:34      -0700   Siebel Universal Agent  64.181.17
TRACE_AREA_OM   TRACE_INFO      2       0       2007-02-13 17:17:28.658 axapp.cpp(236)CSS
TRACE_AREA_OM   TRACE_INFO      2       0       2007-02-13 17:17:28.668 axapp.cpp(10282)C
TRACE_AREA_JAVA TRACE_INFO      2       0       2007-02-13 17:17:28.878 coapp.cpp(4953)JS
TRACE_AREA_BRWS TRACE_INFO      2       0       2007-02-13 17:17:28.878 axcmdmgr.cpp(2470
TRACE_AREA_OM   TRACE_INFO      2       0       2007-02-13 17:17:28.898 axapp.cpp(8960)**
TRACE_AREA_BRWS TRACE_INFO      2       0       2007-02-13 17:17:28.908 axcmdmgr.cpp(2383
TRACE_AREA_REQ  TRACE_INFO      2       0       2007-02-13 17:17:29.018 rpcconnect.cpp(24
TRACE_AREA_REQ  TRACE_DETAIL    3       0       2007-02-13 17:17:29.018 rpcconnect.cpp(28
TRACE_AREA_REQ  TRACE_INFO      2       0       2007-02-13 17:17:31.655 rpcconnect.cpp(48
TRACE_AREA_OM   TRACE_INFO      2       0       2007-02-13 17:17:31.715 axapp.cpp(9534)**
TRACE_AREA_BRWS TRACE_INFO      2       0       2007-02-13 17:17:32.205 axcmdmgr.cpp(2659
TRACE_AREA_BRWS TRACE_INFO      2       0       2007-02-13 17:17:32.235 axcmdmgr.cpp(2545
TRACE_AREA_BRWS TRACE_INFO      2       0       2007-02-13 17:17:32.245 axcmdmgr.cpp(2545
TRACE_AREA_BRWS TRACE_INFO      2       0       2007-02-13 17:17:32.245 axcmdmgr.cpp(2688
TRACE_AREA_BRWS TRACE_INFO      2       0       2007-02-13 17:17:32.245 axcmdmgr.cpp(2470
TRACE_AREA_BRWS TRACE_INFO      2       0       2007-02-13 17:17:32.245 axcmdmgr.cpp(2470
TRACE_AREA_BRWS TRACE_INFO      2       0       2007-02-13 17:17:32.305 axcmdmgr.cpp(2383
TRACE_AREA_BRST TRACE_INFO      2       0       2007-02-13 17:17:32.305 axcmdmgr.cpp(2969
```

Example Where Real Time Streaming Data is Created



E-commerce Purchases

Example Where Real Time Streaming Data is Created

Weather Events

Example Where Real Time Streaming Data is Created

Weather Events

Utility Service Usage

Example Where Real Time Streaming Data is Created

Geo-Location

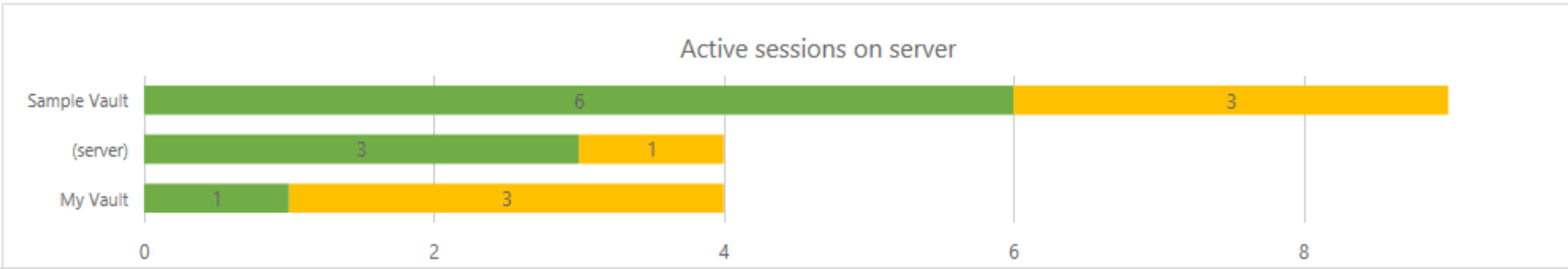
Example Where Real Time Streaming Data is Created (Service Activity)

Refresh
Reset
Hide System Sessions
Export Server Activity...
Import Server Activity...

owing operations for the
st 5 min 8 s.

owing background
ocesses for the last 5 min 3

Active sessions on server



Most active sessions

User	Operation	Total duration	Count	Average	Vault
Mike Taylor	Get value list items	0.415 s	10	0.041 s	Sample Vault
Mike Taylor	Get all value lists	0.234 s	3	0.078 s	Sample Vault
Mike Taylor	Get all property definitions	0.229 s	3	0.076 s	Sample Vault
Mike Taylor	Get all workflows	0.205 s	2	0.102 s	Sample Vault
Mike Taylor	Get all classes	0.150 s	3	0.050 s	Sample Vault
Total		**1.950 s**	**173**	**(0.56 calls per second)**	

Objects modified

(nothing to display)

Views and searches

(nothing to display)

Background processes

Process name	Start time	End time	Duration	Vault
Generate notifications	10/23/2017 10:33:12 AM	10/23/2017 10:33:12 AM	0.007 s	Sample Vault
Generate notifications	10/23/2017 10:32:52 AM	10/23/2017 10:32:52 AM	0.007 s	Sample Vault
Generate notifications	10/23/2017 10:33:55 AM	10/23/2017 10:33:55 AM	0.006 s	Sample Vault
Generate notifications	10/23/2017 10:35:43 AM	10/23/2017 10:35:43 AM	0.006 s	Sample Vault

Data Streaming

- When companies are able to analyze streaming data they receive, they can get real-time insights to understand exactly what is happening at any given point in time.
- This enables better decision-making as well as provide customers with better and more personalized services.
- Nearly every company is or can use streaming data.

Use Cases for

Real-Time & Streaming Data

Use Cases for Real-Time and Streaming Data

- Predictive maintenance:
- Healthcare
- Retail
- Social media
- Finance:
- Energy and power
- Personalization of products and services:
- Transportation and supply-chain
- KPIs

Predictive Maintenance

Predictive maintenance

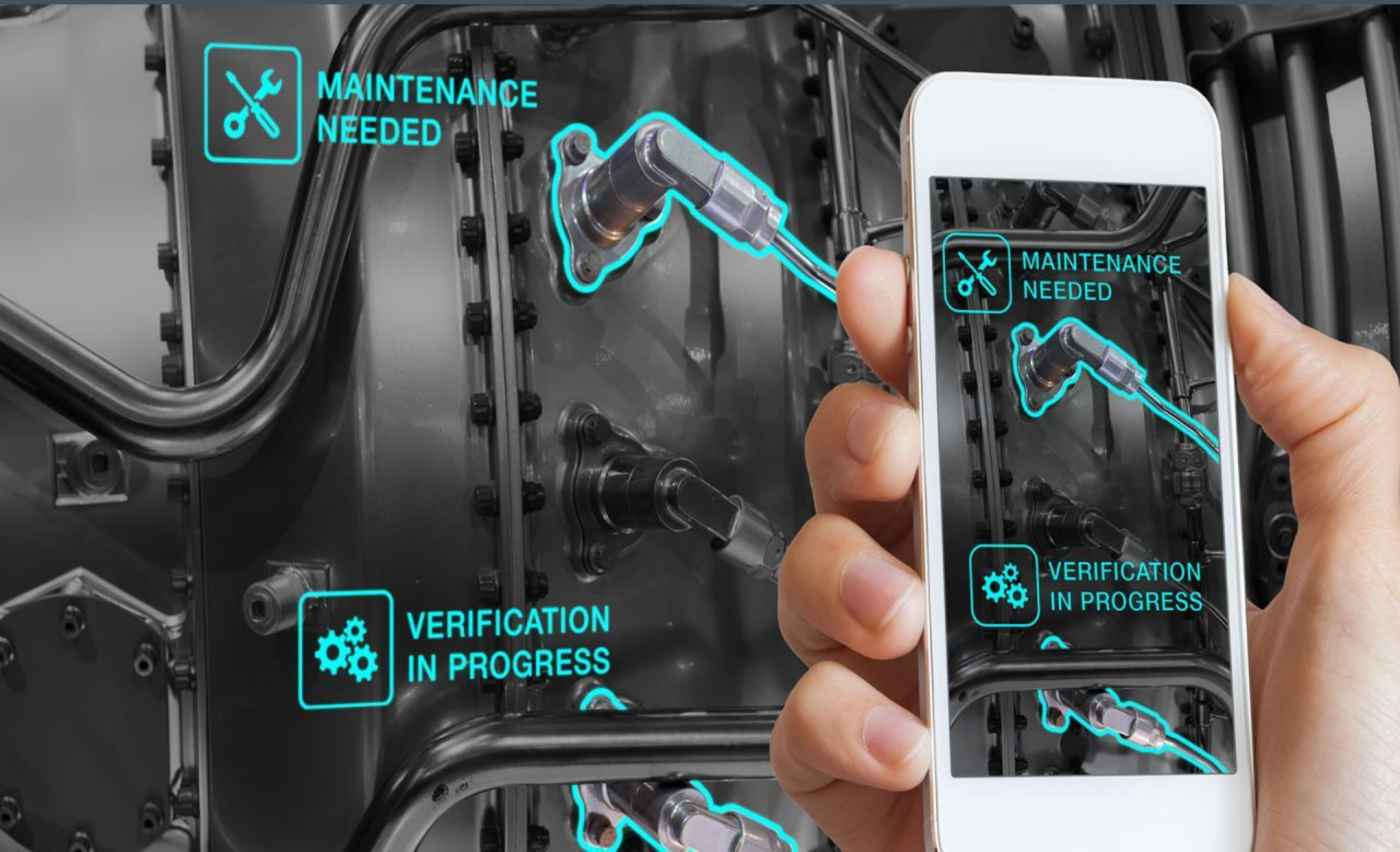
Maintenance Issues

Predictive maintenance

Maintenance Issues

Camera or Sensors

Predictive Maintenance (Monitoring Performance)



Healthcare

HealthCare

Wearable Technology in healthcare

HealthCare

Glucometer

HealthCare

Smart Scales

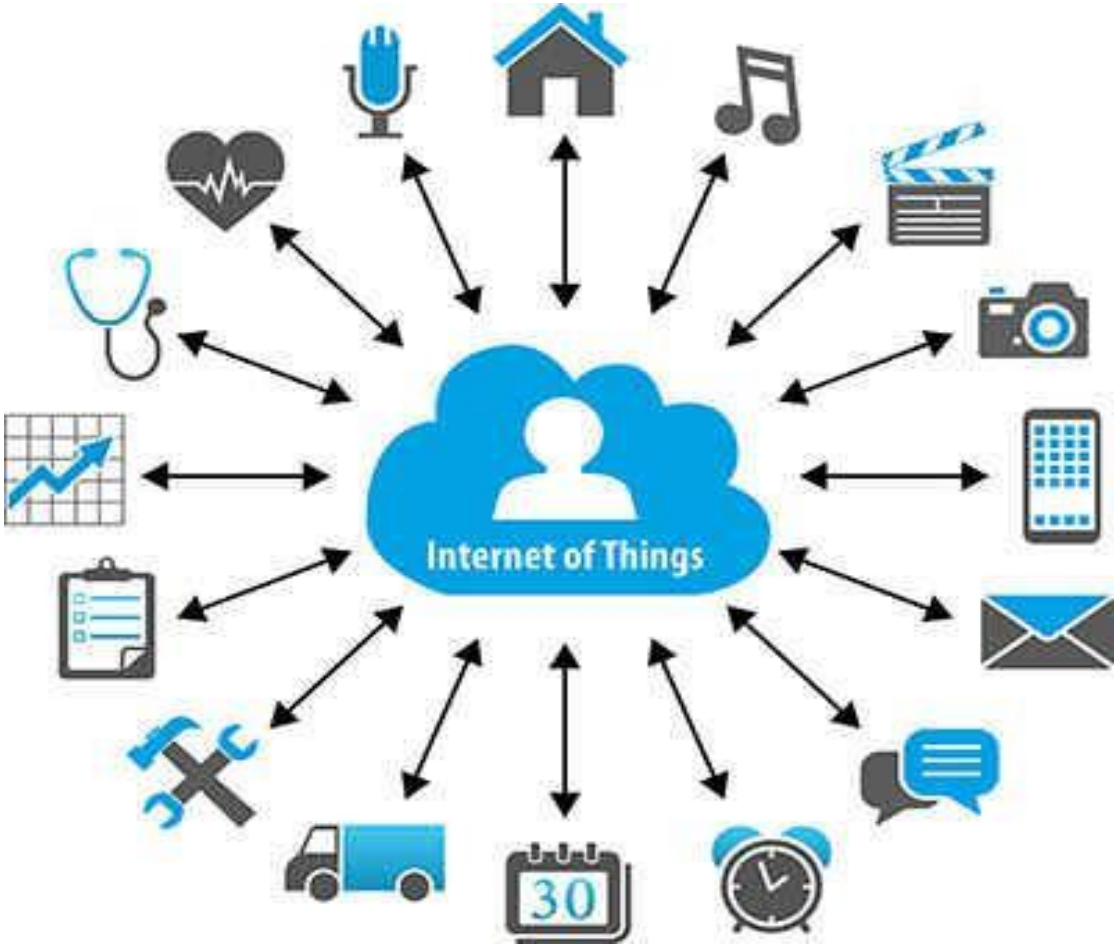
HealthCare

Heart Rate and Blood Pressure Monitor

Retail

Retail

Retail



IOT Sensors

Retail

Retail

Brick and Mortal Stores

Retail

- What are possible with real time data?

Location Based Marketing

THE TREND INSIGHTS®

Retail

- What are possible with real time data?

Location Based Marketing

THE TREND INSIGHTS®

Trend Insight

Retail

- What are possible with real time data?

Location Based Marketing

Trend Insight

Improvement to Operational efficiency

Product Movement

Product Freshness

Social media

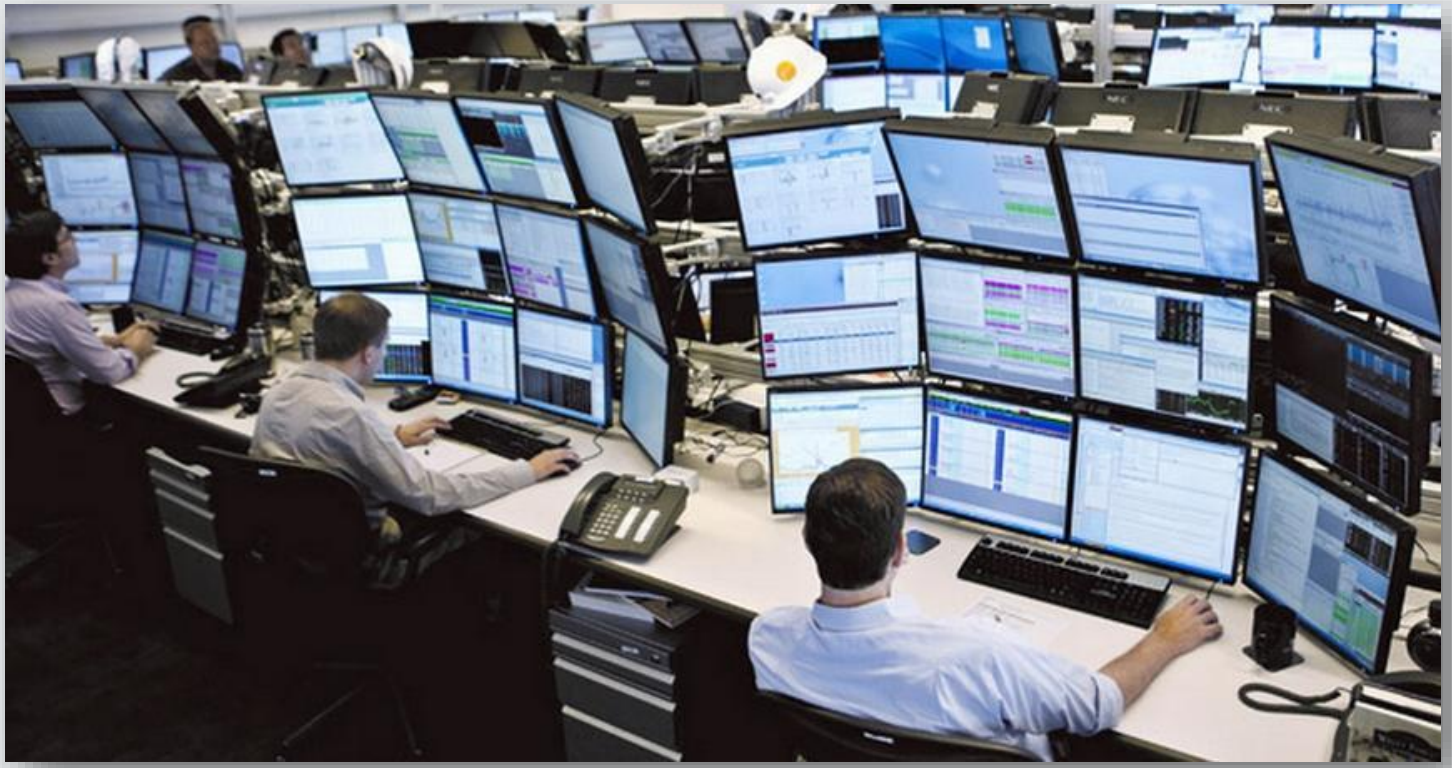
Social media

Social media

Social media bullying

Finance

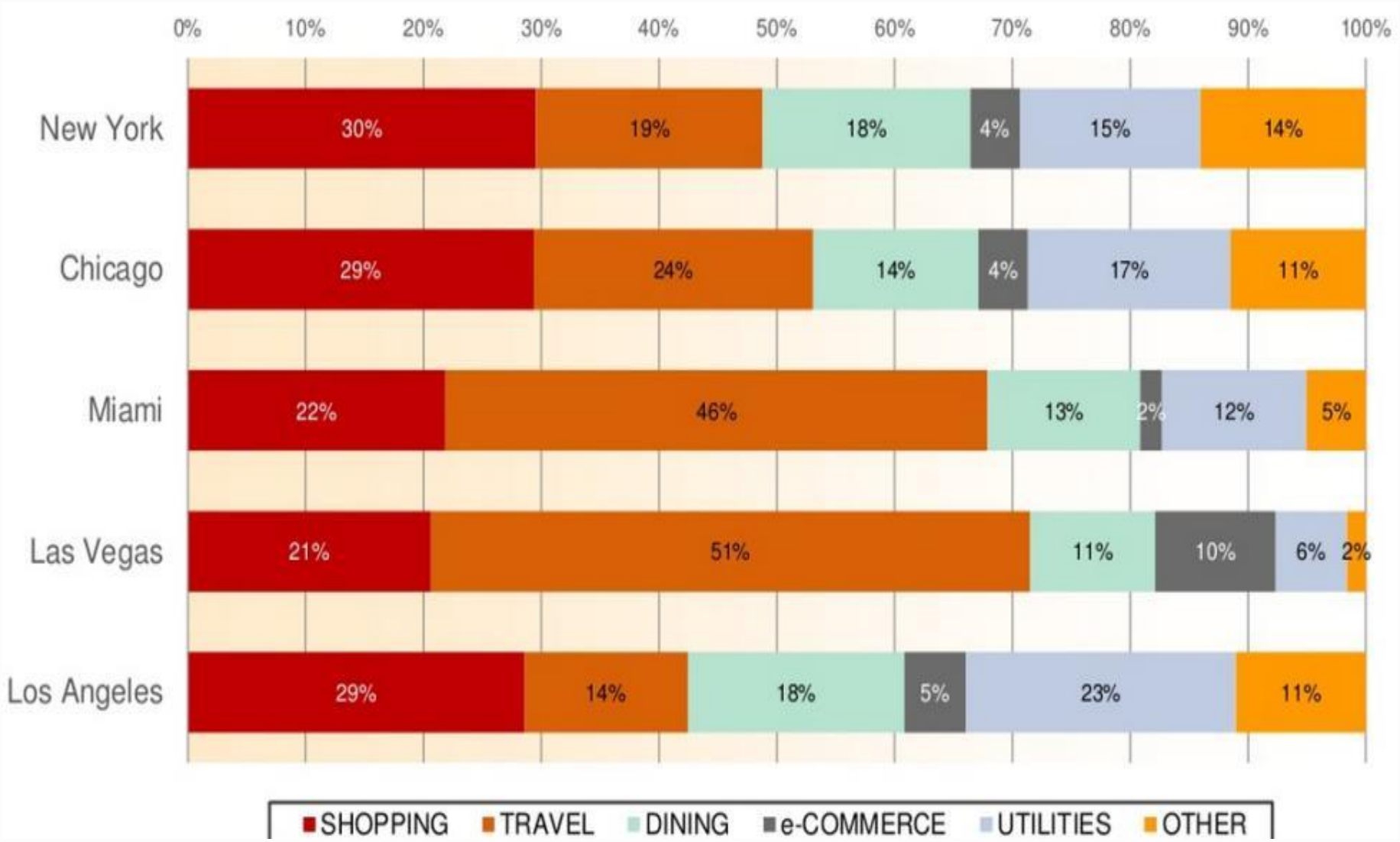
Finance



Trading floors

Finance

Finance (Identify Fraudulent Merchants)



Finance

Energy and Power

Energy and Power

Fossil Fuels

SUSTAINABLE
POWER SYSTEMS

Personalization of Products and Services

Personalization of Products and Services

Streaming Data

Personalization of Products and Services

Online News Publication

Transportation and Supply-chain:

Transportation and Supply-chain:

Internet of Trains

KPIs

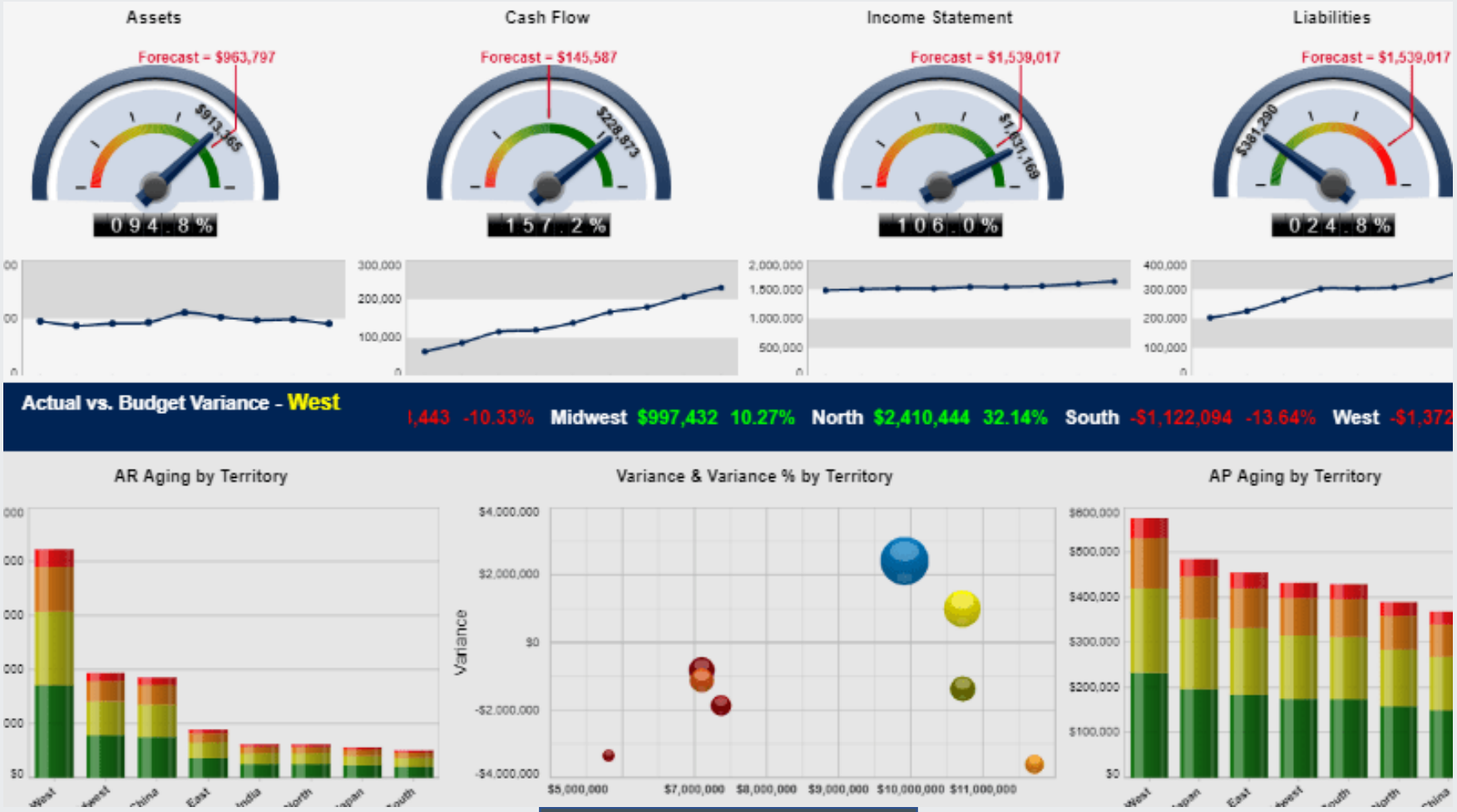
KPIs

Real Time KPIs

KPIs

Financial

KPIs



Financial

Data Lake

What is a Data Lake?

Why do You Need a Data Lake?

Organizations that successfully generate business value from their data, will outperform their peers.

Why do You Need a Data Lake?

This helped them to identify, and act upon opportunities for business growth faster by attracting and retaining customers, boosting productivity, proactively maintaining devices, and making informed decisions.

Why do You Need a Data Lake?

This helped them to identify, and act upon opportunities for business growth faster by attracting and retaining customers, boosting productivity, proactively maintaining devices, and making informed decisions.

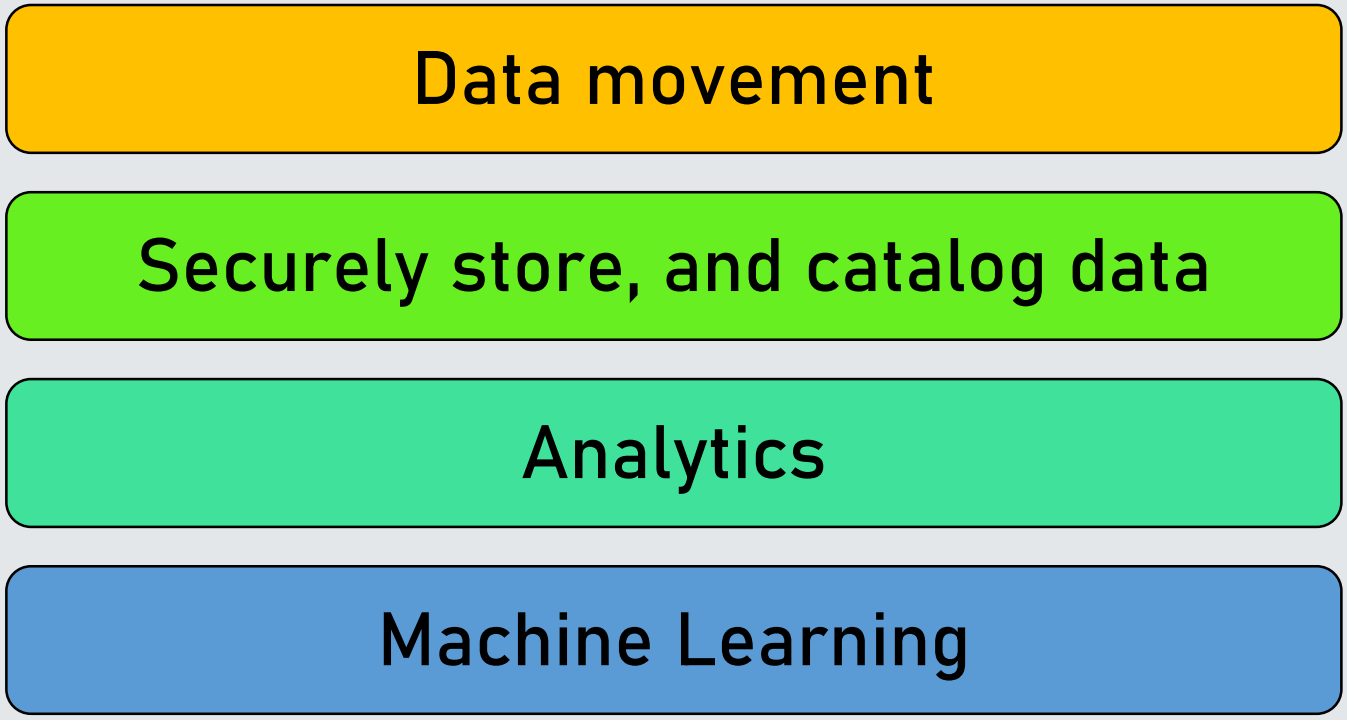
Data Lakes compared to Data Warehouses

Characteristics	Data Lakes	Datawarehouse
Data	Relational from transactional systems, operational databases, and line of business applications	Non-relational and relational from IoT devices, web sites, mobile apps, social media, and corporate applications
Schema	Designed prior to the DW implementation (schema-on-write)	Written at the time of analysis (schema-on-read)
Price/Performance	Fastest query results using higher cost storage	Query results getting faster using low-cost storage

Data Lakes compared to Data Warehouses

Characteristics	Data Lakes	Datawarehouse
Data Quality	Highly curated data that serves as the central version of the truth	Any data that may or may not be curated (ie. raw data)
Users	Business analysts	Data scientists, Data developers, and Business analysts (using curated data)
Analytics	Batch reporting, BI and visualizations	Machine Learning, Predictive analytics, data discovery and profiling

The Essential Elements of a Data Lake and Analytics Solution



The Value of a Data Lake

Improved customer interactions

Improve R&D innovation choices

Increase operational efficiencies

The Challenges of Data Lake

Data Lake Challenges

Manual processes requiring hand-coding and reliance on command-line tools

Hard to find data and its lineage for data discovery and exploration

Coupling of ingestion and processing drives architecture decisions

Operationalizing processes for production and to maintain SLAs

Ensuring data is in canonical forms with a shared schema usable by others

Coding or filing tickets often required to perform new ingestion and processing tasks

Multiple architectures and technologies used by different teams on different clusters

Guaranteeing compliance in a system that is designed for schema-on-read and raw data

Sharing infrastructure in a multi-tenant environment without low-level QoS support


That's all for now...