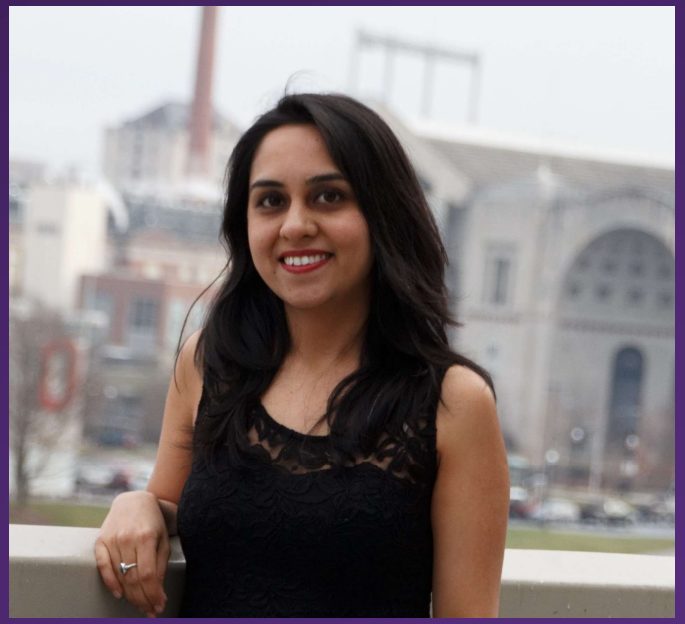
Unlocking the Power of Data Observability and Monitoring

Hello!

I am Roma Nawani

I am here because I love everything about data.

You can find me at LinkedIn @RomaNawani



Copy Of Slides

Will be uploaded on Github.com

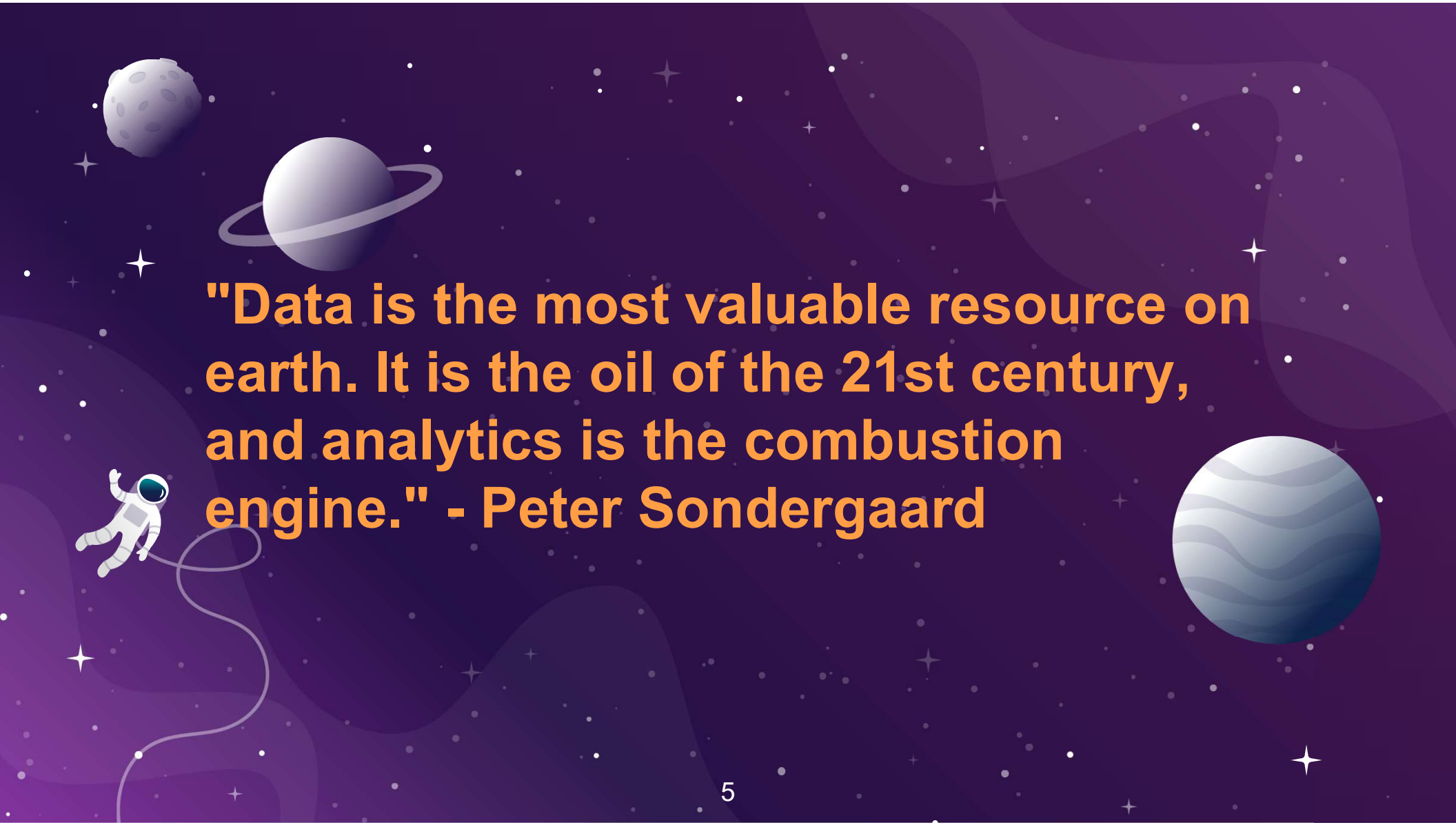
Link : https://github.com/romanawani/codemash2023

Session Synopsis

""When will I receive my data?" "What is causing delays?". If you have been receiving these messages from the business stakeholders, you are not alone.
Many data teams spend over 30% of their time addressing data pipeline issues, troubleshooting problems, missing SLAs, and dealing with incorrect and imprecise data. As we rely more on hybrid infrastructure and systems that become increasingly complex and distributed, this trend is likely to become more pronounced in the future.

As Data Observability gains importance in the data stack, it aids engineers and analysts in minimizing the manual effort required to identify issues caused by incorrect data, code, and operational problems. However, how does the practical implementation of data observability and monitoring appear in reality?

This session will provide an overview of the key components of Data Observability and cover best practices for data teams looking to achieve comprehensive visibility into their data at scale.


"Data is the most valuable resource on earth. It is the oil of the 21st century, and analytics is the combustion engine." - Peter Sondergaard

Agenda

1. Intelligent Digital Data Enterprise Journey
2. Modern Data Architecture
3. What is Data Observability?
4. Data Observability Practices
5. Data Observability vs Monitoring
6. Data Observability Tools
7. Summary
8. Appendix

1. Digital Data Enterprise

Transformation Journey

Intelligent Digital Data Enterprise

- Cloud is the new normal
- Strong foundation for Data
- AI/ML for Intelligent decision

Data as a Platform

Data Pipelines

Pipelines Complexity

Recent advancements in the data field have been influenced by technologies such as the Internet of Things (IoT), serverless computing, hybrid cloud, artificial intelligence (AI), and machine learning (ML).

Data Volume

Data Volume

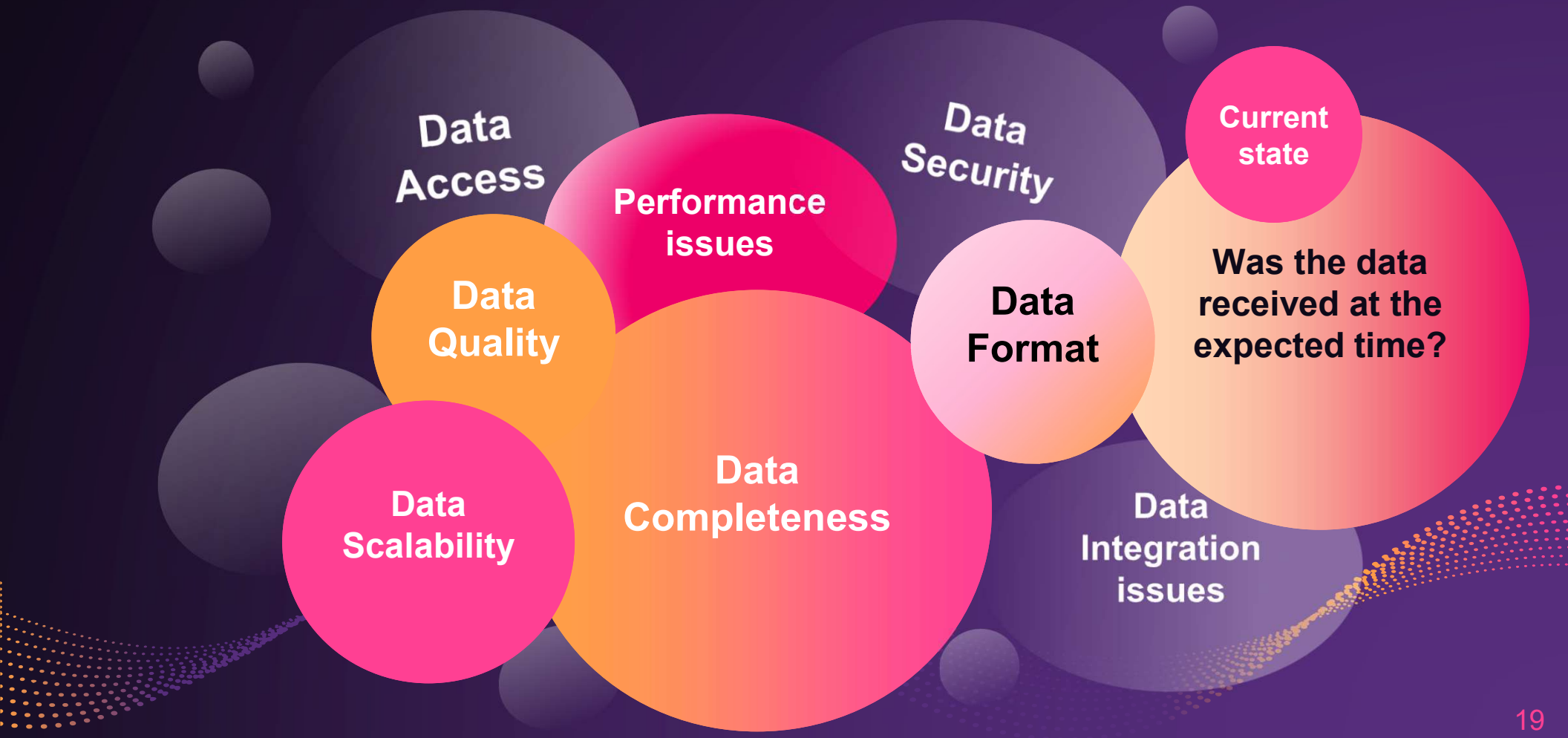
Data Pipelines

2. Modern Data Architecture

Digital Data Enterprise Journey

- More External Data Sources
- Multiple Tools and Platforms
- Complicated Transformations
- Hybrid Data Architecture
- Real Time/Near Real Time processing
- Complex Data Pipelines

Common Problems



What do our Stakeholders need ?

Data Developer

Production Engineer

Data Team Leader

Business Users

- Developers: Data Developers/Data Engineers, BI Developers
- Production Operations Team
- Data Team Leaders / App Owners / Product Managers / Delivery Leads/Data Asset Owner
- Business Users / Business Customers, Data Scientists, Data Analysts

What do our Stakeholders need ?

What is current status of the data and pipelines?

What is the source of the delays in the delivery of the data?

Data Developer

Production Engineer

Data Team Leader

Business Users

3. Data Observability

Data Observability

Data observability is an organization’s ability to fully understand the health of the data in their systems and Data pipelines.

Pillars of Observability for Data Systems

Tracking the path of data as it flows through various pipelines. Identify critical path for data pipelines.

Measuring Quantifiable aspects of the data system such as resource utilization throughput, latency, and error rates.

Capturing and storing log data generated by the system , which can be used for debugging, analysis and alerting.

Collecting and storing data about significant changes in the system

Why Data Observability is important?

Visibility

Provide visibility to the business customer on data delivery progress

Trust

Establishing trust in the data so that business can take more confident data-driven decisions

Data Volume

Monitor spike in the data volume as high or low volume could be due to broken pipeline .

Data Completeness

Increasing the usefulness and completeness of the data

Data Timeliness

Ensuring Data is delivered in a timely manner within defined SLA for business decision making

Data Delivery

Track data is delivered to defined destination

4. Data Observability Practices

Data Observability Practices

Logging

Tracing

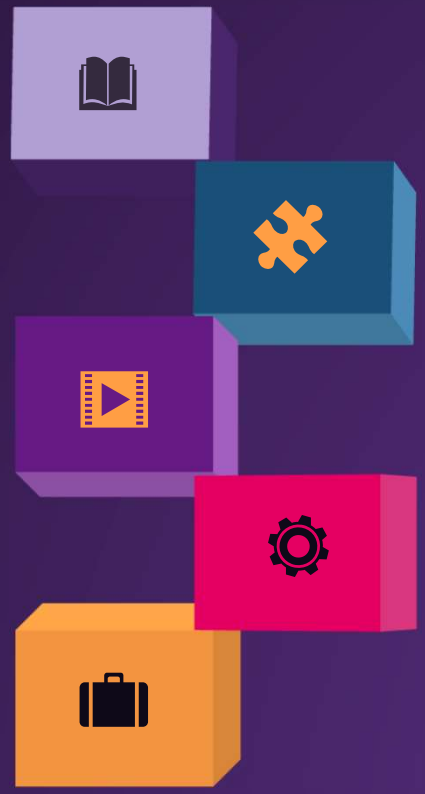
Metrics

Alerting

Correlation and Root Cause Analysis

Visualization

Observability Building Blocks



Data Access and Usage

How are teams using data? What critical processes are involved ?

Data Trends

Are there any important changes in the data that I should know about ?

Data Sanity

Is the data valid and complete ? Are there any errors in the data itself ?

Pipeline Latency

Is data arriving within defined SLA ? Are you meeting your SLAs?

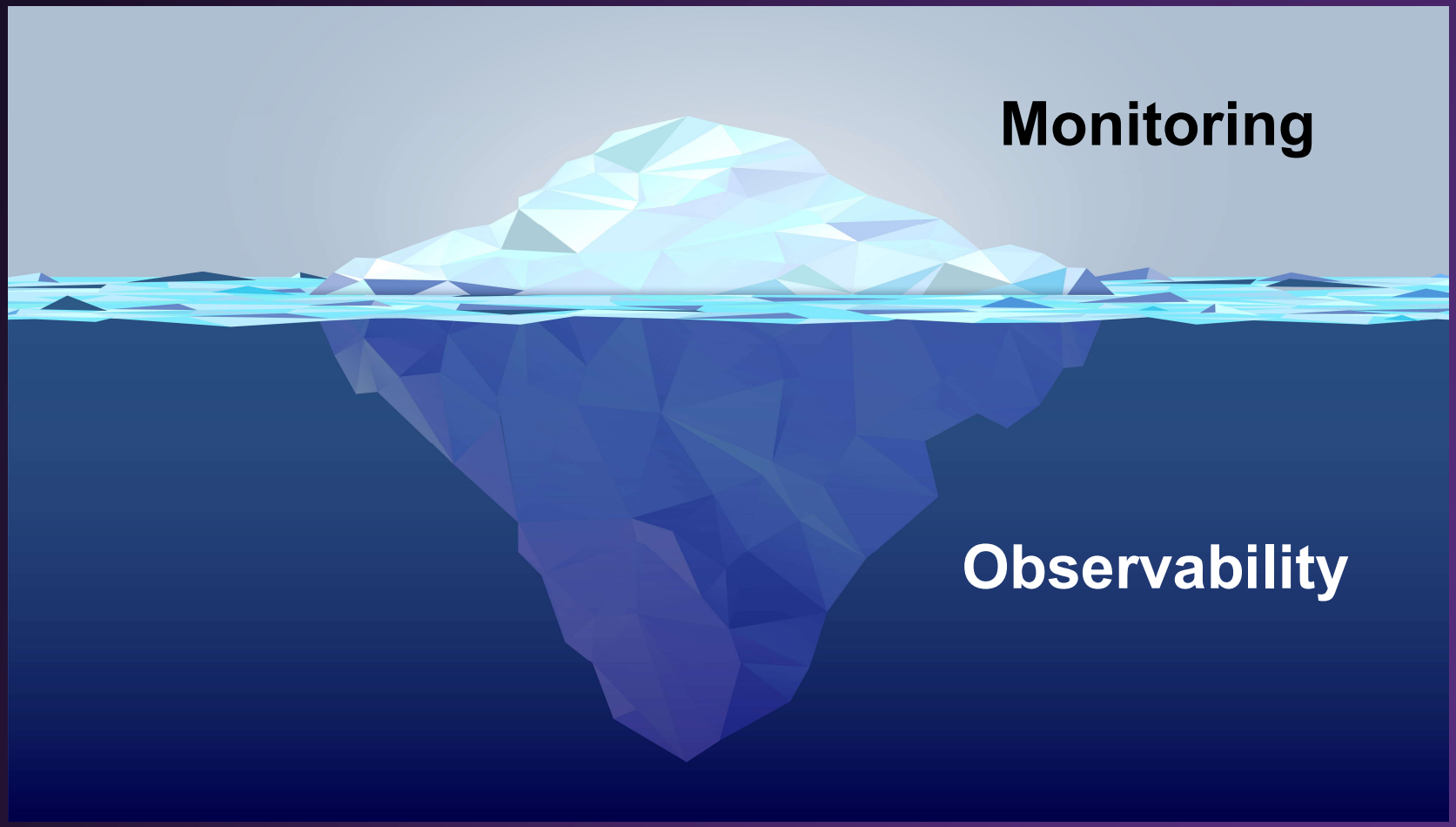
Monitoring Data Pipelines

Is the data flowing through my pipelines?

Data Observability
&
Enterprise Strategy

Having accurate, reliable, and trustworthy data leads to more efficient data teams, satisfied customers, and increased utilization of data

5. Data Observability vs Monitoring



Proactive vs Reactive Alerting

Proactive:

- Alert when disk consumption reached certain percentage.
- Alert when table size is growing rapidly
- Alert when database table is approaching storage limit before it runs out of space
- Track volume spikes and drops also helps us identify a lot of "human invisible" issues.

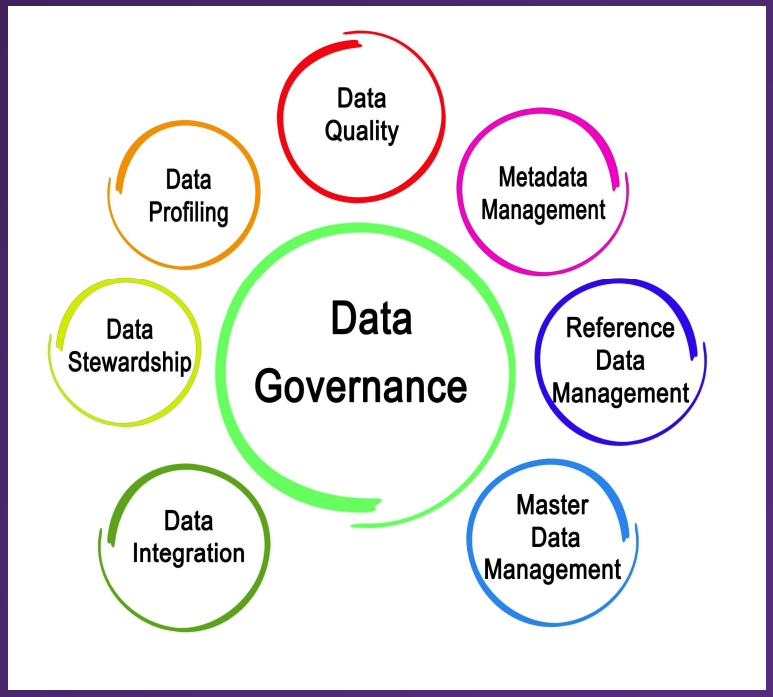
Reactive:

- Alert when disk is out of space
- Alert when data replication fails
- Alert when query takes long time to complete.
- Threshold based alerts ; “alert me when the SLA has been breached”.

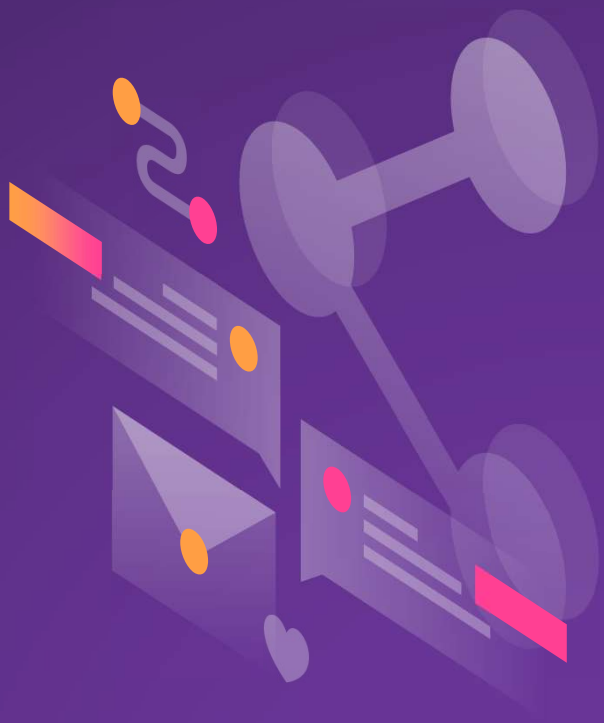
Data Observability vs Data Governance

Data Observability

Data Governance



6. Data Observability Tools



Data Observability Platform Features

End to End Visibility

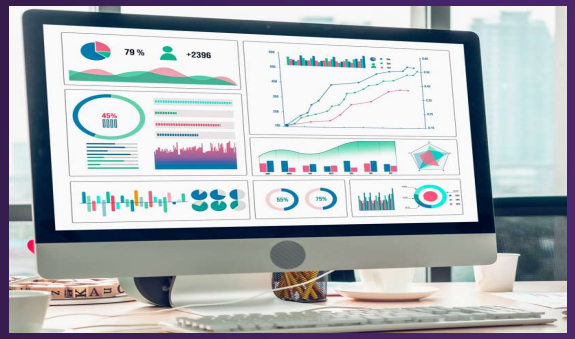
Real Time Monitoring

Anomaly Detection

Security

Debugging

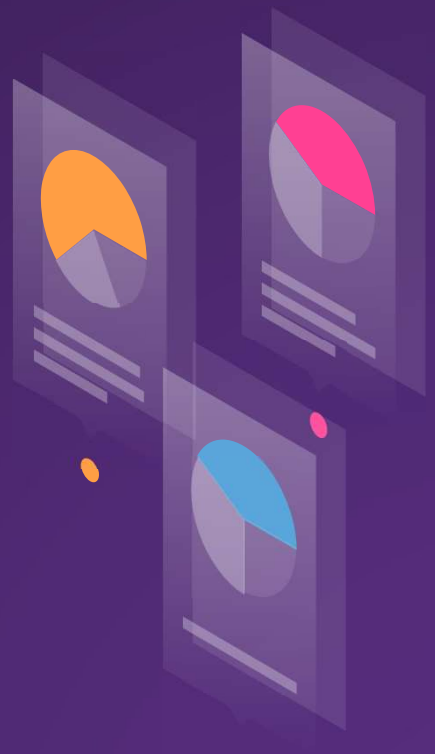
Customizable dashboards



Log aggregation and visualization

Application Documentation

- System / Application
- Source
- Destination
- Log format
- Define Event type
- Define Thresholds
- Customers/Consumers
- Frequency
- SLA



New Relic One

New Relic Features:

- Real-time monitoring and alerting
- Distributed tracing
- Log management
- Event data
- Dashboards and analytics

New Relic

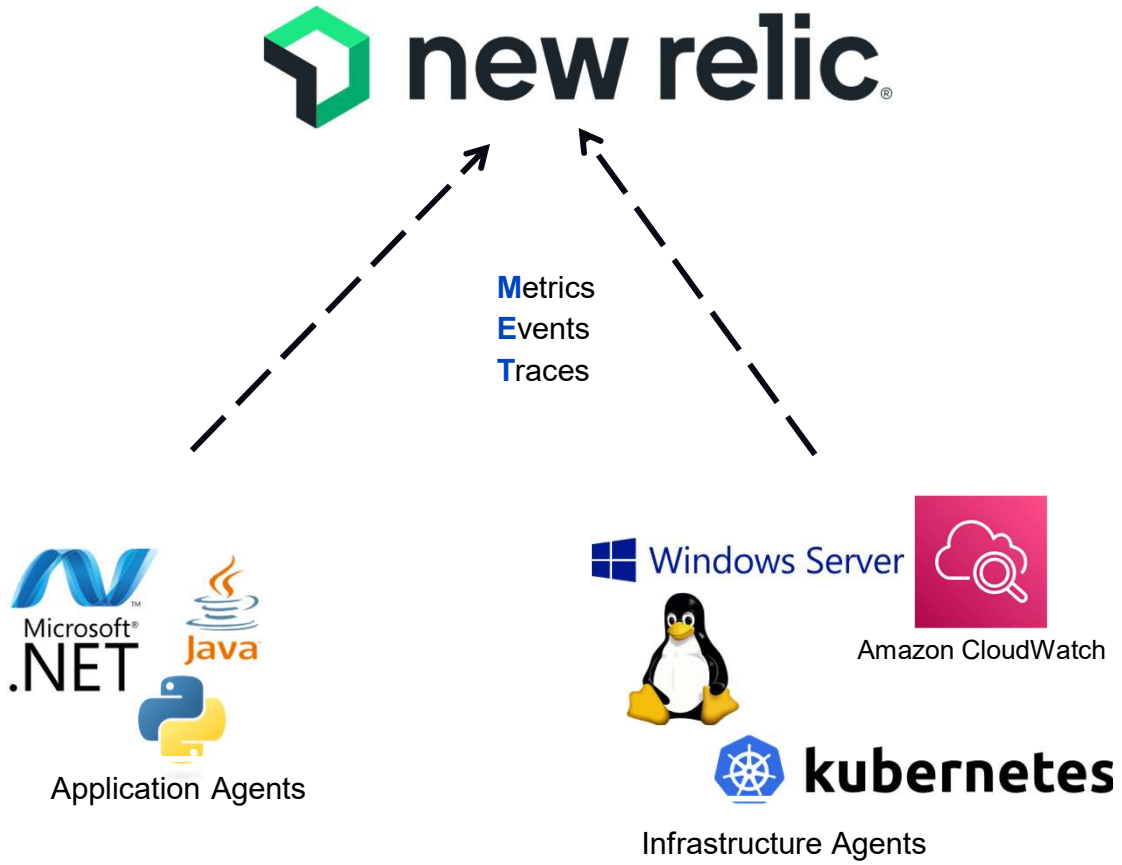
An observability platform that provides visibility into the health and performance of your technology stack.

How?

- By collecting infrastructure and application data (M.E.L.T.)
- By providing a query language, NRQL to analyze telemetry data. (Dashboards, charts)
- By providing tools to proactively detect anomalies and potential problems before they become critical.

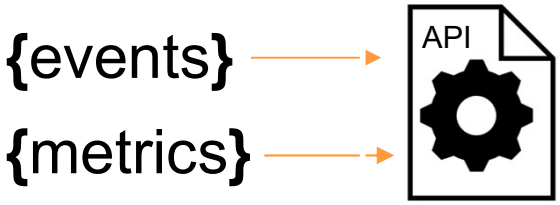
Goal: Identify weaknesses and make improvements. Build better software!

NOT just another alerting and synthetics tool.



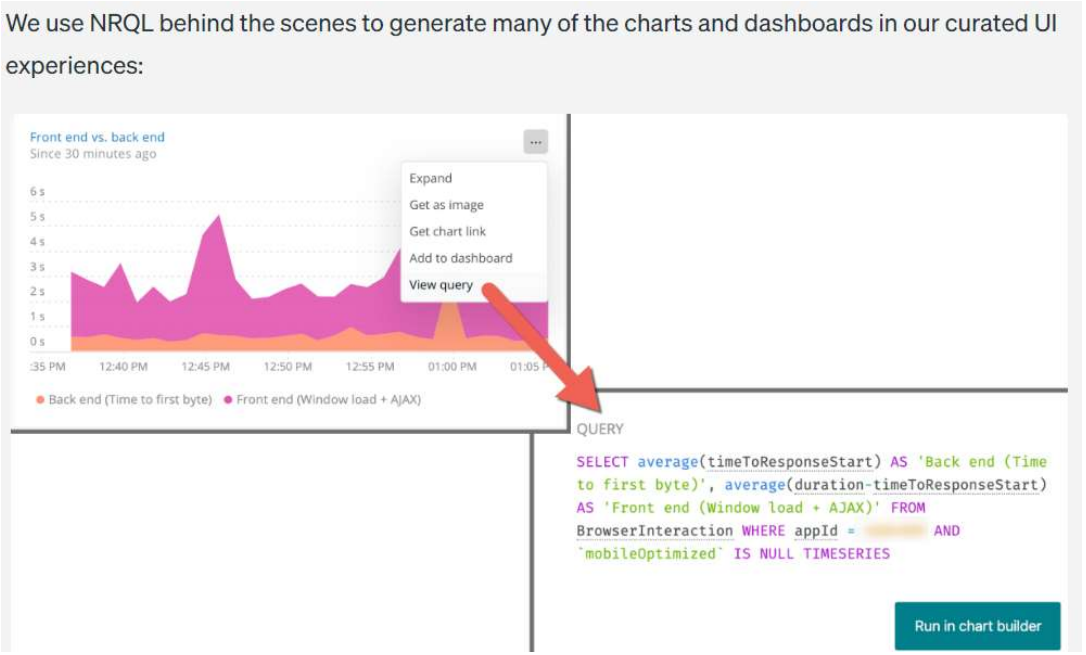
What if I cannot install an agent?

You can send custom data by using New Relic's Metrics and Events API.



https://docs.newrelic.com/docs/apis/intro-apis/introduction-new-relic-apis/

Example: Send metadata such as Job start time, job end time, job status, And data volume to New Relic.



New Relic Use Cases

Database Monitoring: Collect performance metrics from databases and provide detailed views of database performance, including slow queries, errors and connection count.

Data Storage Monitoring: Monitor for data size, read/write operations, and errors, to understand how your data storage is performing and how it impacts your systems. Compare average data writes to a file system between two periods.

Events Monitoring: Gain insights into what's happening inside your data systems with a detailed view of the events, from user sessions, to background jobs, to external API calls.

Infrastructure Monitoring: View resource consumption of a k8s container, CPU, memory on servers

Data pipeline monitoring: Detailed metrics such as incoming and outgoing data rate, number of messages in queue, and record processing time

Acceldata

Acceldata Products:

- **DATA OBSERVABILITY CLOUD**
- **PULSE**
- **TORCH**

Acceldata Tool provides:

- Real-time monitoring of data pipelines, data lakes, and data warehouses
- Root cause analysis of performance issues
- Predictive analytics and alerting
- Drill down from high-level views to specific data points
- Auto-discovery of data flows
- Multi-cloud support, can be deployed in any cloud environment such as AWS,Azure,GCP

MonteCarlo

Provides End-to-end solution for data stack that monitors and alerts for data issues across data warehouses, data lakes, ETL, and data lakehouse.

Monte Carlo comes with several features like data catalogs, automated alerting, and observability on several criteria out of the box

Splunk

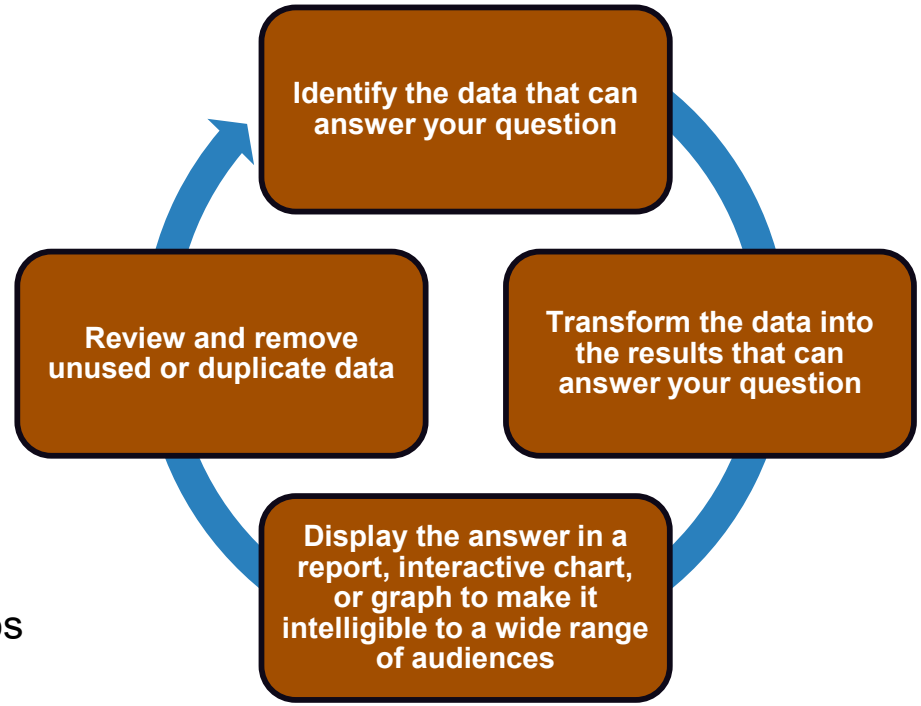
Splunk Features:

- Real-time Monitoring
- Log Management
- Security Information and Event management
- Compliance
- Machine learning

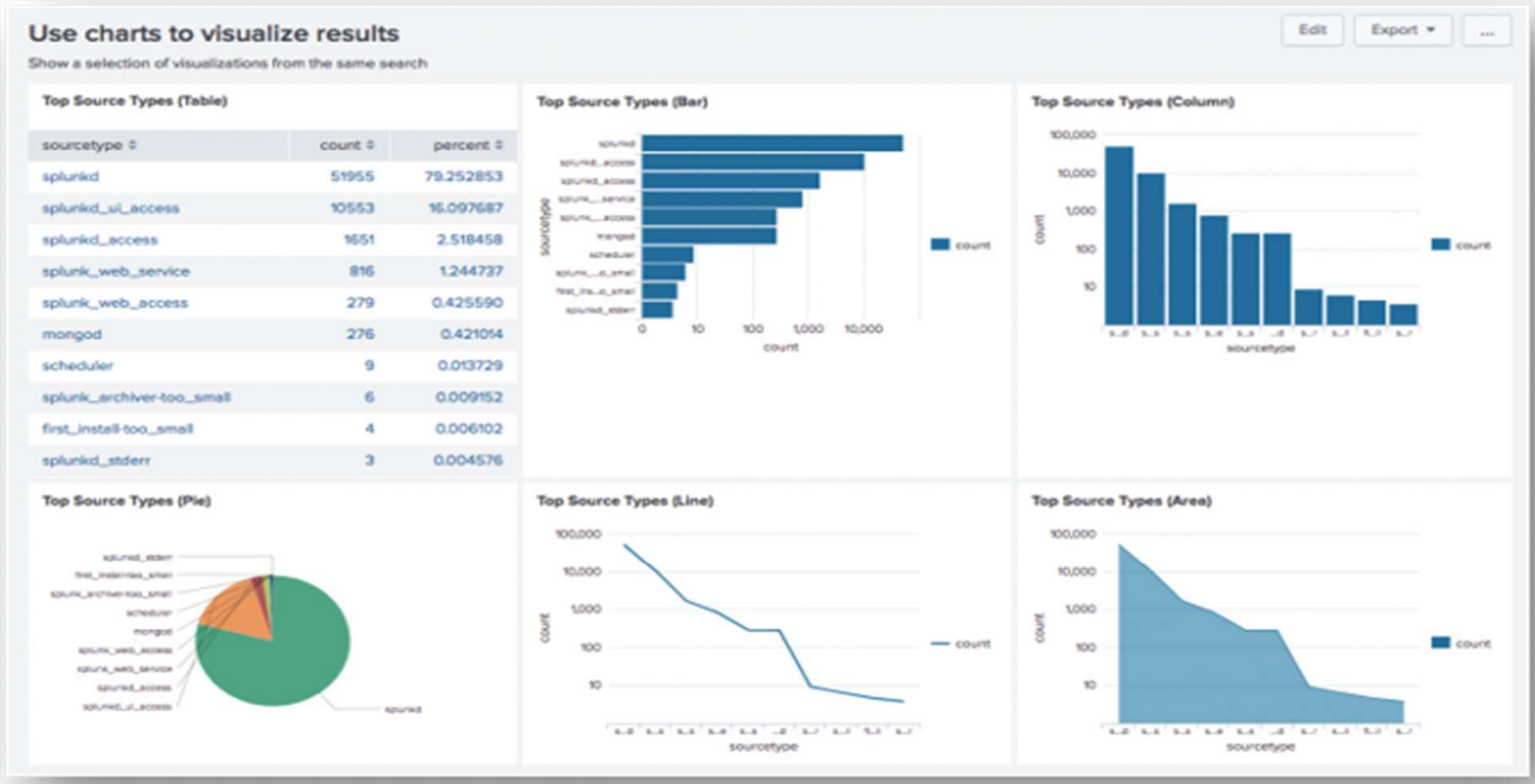
Splunk Usages and Workflow

- Users' interaction with an App
- Service goes down
- Number of users for an application
- Find trends in failed chatbot interactions
- Comparing A/B testing results
- Execution time of automated tests
- Access Denies
- System Logs
- Application Logs

https://education.splunk.com/category/use-case-videos



Real Time Dashboards



Dynatrace

Provides real-time insights into the performance and behavior of software systems and services.

Dynatrace Features:

- Real-time Monitoring and alerting
- Automatic Root Cause Analysis
- AI/ML to analyze data and detect issues
- Automated Problem resolution
- Cloud-native support

Datadog

Datadog is a cloud-based monitoring and observability platform .

Datadog Platform Features:

- Real-time Monitoring
- Application Performance management
- Log management
- Distributed tracing
- Anomaly detection

Tools Comparison

New Relic

- Real-time Integration
- Offers a wide range of integrations and plugins
- Provides detailed performance data and troubleshooting

Splunk

- Platform for searching, analyzing, and visualizing machine-generated data
- Wide range of use cases, including security, IT operations, and business analysis
- Strong focus on data collection and analysis

Dynatrace

- Provides real-time performance monitoring and troubleshooting
- Offers a wide range of integrations and plugins
- Provides detailed performance data and troubleshooting

Datadog

- Provides real-time performance monitoring and troubleshooting
- Offers a wide range of integrations and plugins
- Has a focus on providing detailed performance data and troubleshooting

Custom Solution

Process Date: 1/11/2023

Current Time: 1/12/2023 08:30

Source	Target	Defined SLA	Start Timestamp	Completion Timestamp	SLA Status	Pipeline Status
Application A	Application B	22:00	1/11/2023 20:00	1/11/2023 21:45:09	SLA MET	Complete
Application A	Application C	22:00	1/11/2023 20:00	1/11/2023 22:00:09	SLA Breached	Complete
Application B	Application C	8:30	1/12/2023 5:00	1/12/2023 7:30:00	SLA MET	Complete
Application D	Application C	7:00	1/12/2023 5:00	1/12/2023 6:30:08	SLA Breached	Complete
Application C	Application E	8:00	1/12/2023 5:00	1/12/2023 8:30:08	SLA Breached	Complete
Application W	Application Z	6:00	1/12/2023 6:34		SLA Breached	Inprogress
Application F	Application C	6:00	1/12/2023 2:34	1/12/2023 5:58:08	SLA MET	Complete
Application F	Application Z	8:00			SLA Breached	Not Started

Custom Dashboard

Trend Last 12 days

Source	Target	Defined SLA	11-Jan-23	10-Jan-23	9-Jan-23	8-Jan-23	7-Jan-23	6-Jan-23	5-Jan-23	4-Jan-23	3-Jan-23
Application A	Application B	22:00	✓	✓	✓	✓	✓	✓	✓	✓	✓
Application A	Application C	22:00	✗	✓	✓	✓	✓	✓	✓	✓	✓
Application B	Application C	8:30	✓	✓	✗	✓	✓	✓	✓	✓	✗
Application D	Application C	7:00	✗	✗	✗	✓	✗	✓	✓	✓	✗
Application C	Application E	8:00	✗	✗	✗	✓	✓	✗	✓	✓	✗
Application W	Application Z	6:00	✗	✗	✓	✓	✓	✓	✓	✓	✓
Application F	Application C	6:00	✗	✗	✗	✗	✗	✓	✗	✗	✗
Application F	Application Z	8:00	✗	✓	✗	✗	✗	✗	✗	✓	✗

7. Summary

Evaluating Data Observability Platform

- ✓ Can the tool provide insights into the behavior of your microservices/systems/data pipelines over time and in different situations?
- ✓ Is the tool able to gather and display all necessary data in a single location?
- ✓ At what speed does the tool notify you of any issues? Is it able to do so in real-time?
- ✓ Does it provide sufficient context and information to understand the problem?
- ✓ What methods does the tool use for collecting and storing metrics, logs, and traces?
- ✓ How does it ensure the integrity and accuracy of data collected?
- ✓ Does the tool provide context on incidents — what went wrong, which services were affected, and what was the impact on business or customer experience?
- ✓ Does it connect smoothly to your current data infrastructure, tools and platform?
- ✓ Are there any reviews or testimonials for the tool on industry-standard portals such as Gartner, G2, or Capterra?

Use Case list for Data Observability:

- Capture metadata about data pipelines such as start time/end time
- Log Monitoring
- Pipeline Latency
- Data Volume Monitoring
- Infrastructure Monitoring
- Custom Metadata Collection
- Batch Monitoring
- Custom Dashboards for Business and Operation Support
- Monitor Real time data behavior
- Monitor Security Events

Data Observability vs DataOps

Data observability is as essential to DataOps as Observability is to DevOps.

Data Observability Best Practices

- Better Data Observability starts at your sources, not your target.
- Do not monitor everything
- Put alerts for only critical touchpoints
- Do not store all logs and data (define cleanup strategy)
- Create custom graphs according to the customer needs

Appendix

Terminology

- **Data Scientists:** Data scientists are responsible for performing statistical analysis using machine learning and artificial intelligence on collected data in order to gain insights and form new hypotheses.
- **Data Lake:** A data lake is a storage repository that holds a vast amount of raw data in its native format. With the modern cloud implementation data can be stored in Raw and Optimized formats. This optimization is not based on data transformation, but reorganization of the same data into file structures that promote the fit for purpose of the intended query.
- **Data Warehouse:** Contains historical and integrated data that has been standardized to support broad set of use cases
- **Data Mart:** Contains subject oriented data repository that serves specific line of business. This data in data mart is slightly curated and is sources from either data lake or Datawarehouse.
- **Data Lakehouse:** Combines the best of Data Warehouse and Data Lake. This is an open-source architecture implementing data structures and features of Datawarehouse directly on low-cost cloud storage.
- **DataOps:** DataOps, short for "Data Operations," is a set of practices and tools that are used to improve the speed, quality, and reliability of data in an organization. The goal of DataOps is to improve the collaboration, automation, and measurement of data processes in order to make data more accessible, accurate, and valuable to the organization.

Terminology

- **DevOps**: **DevOps** is a set of practices that combine software development (*Dev*) and IT operations (*Ops*). It aims to shorten the systems development life cycle and provide continuous delivery with high software quality. DevOps is complementary with Agile software development; several DevOps aspects came from Agile methodology.
- **Data Mesh**: This is based on the idea of breaking monolithic data architectures into many small, autonomous data services, each of which is responsible for a specific domain of data.
- **Data Fabric**: is a design concept to provide a single, unified view of an organization's data, making it more accessible, usable, and valuable for the organization.

Index of Resources

Tools	Links
Big Panda	https://a.bigpanda.io/
DataOps	https://www.gartner.com/en/information-technology/glossary/dataops
Demystifying MELT	https://itbrief.com.au/story/demystifying-m-e-l-t-the-key-data-for-business-observability
Introduction to New Relic APIs	https://docs.newrelic.com/docs/apis/intro-apis/introduction-new-relic-apis/
Splunk	12 Days of Splunk Use Cases | Splunk
NewRelic Agents	https://docs.newrelic.com/docs/new-relic-solutions/new-relic-one/install-configure/install-new-relic/
NRQL	https://docs.newrelic.com/docs/query-your-data/nrql-new-relic-query-language/get-started/introduction-nrql-new-relics-query-language/
Infrastructure data	https://docs.newrelic.com/docs/infrastructure/manage-your-data/data-instrumentation/default-infrastructure-monitoring-data

Index of Resources

Tools	Links
BigPanda	https://a.bigpanda.io/
MonteCarlo	https://www.montecarlodata.com/product/data-observability-platform/
Data Kitchen	https://datakitchen.io/
Datadog	https://www.gartner.com/reviews/market/application-performance-monitoring-and-observability/compare/datadog-vs-newrelic
Dynatrace	https://www.gartner.com/reviews/market/application-performance-monitoring-and-observability/compare/dynatrace-vs-newrelic
Acceldata	https://www.acceldata.io/product-videos

QUESTIONS ?

Copy Of Slides

Will be uploaded on Github.com

Link : https://github.com/romanawani/codemash2023

Thanks!

More questions?

You can find me at

- LinkedIn@RomaNawani
- roma.gurnani@gmail.com

THE END