Big Data

Big Data Course

Eng.Mostafa Nabieh

Mostafa Nabieh

وزارة الاتصالات وتكنولوجيا المعلومات

MANNING

وزارة الاتصالات وتقنية المعلومات
MINISTRY OF COMMUNICATIONS AND INFORMATION TECHNOLOGY

AI TALENTS HUB

UDACITY

PLURALSIGHT

Learning Space
YOUR SPACE TO LEARN FUTURE SKILLS

BIG DATA

Mostafa Nabieh

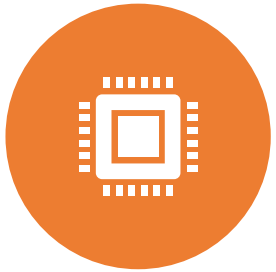
HUAWEI

EPITA
ÉCOLE D'INGÉNIEURS EN INFORMATIQUE

IBM

Mostafa Nabieh

CONTENTS



WHAT IS DATA ENGINEERING?

BIG DATA ECOSYSTEM

BIG DATA LIFECYCLE

CAREER OPPORTUNITIES

Performance Tuning and Troubleshooting

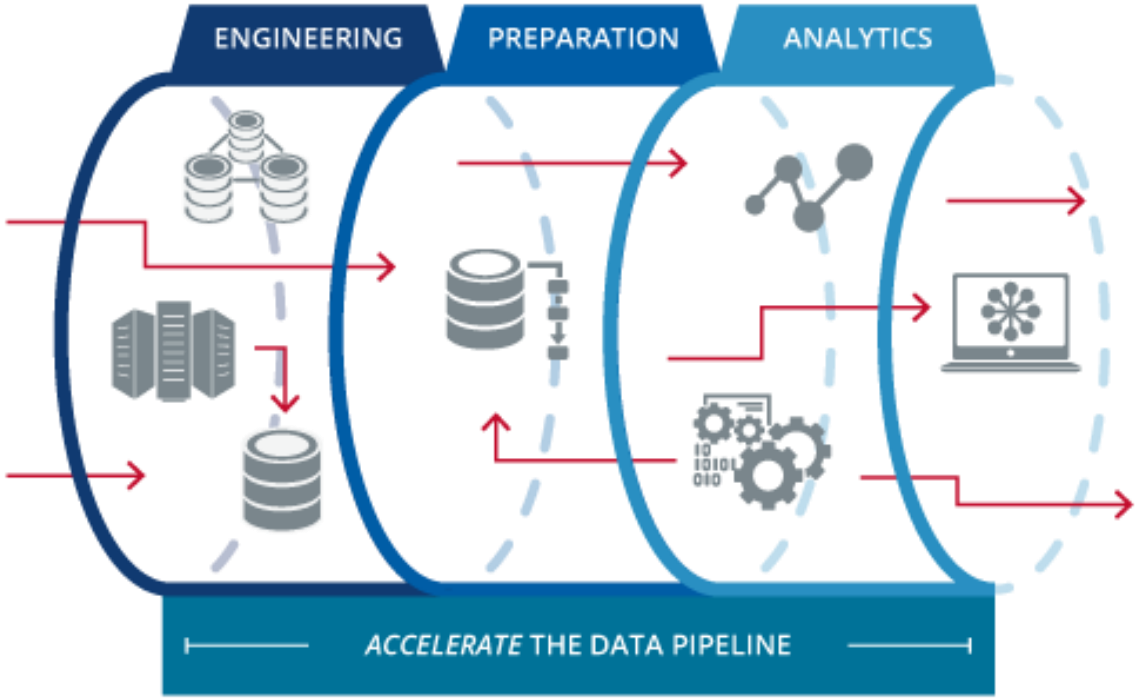
One of the key responsibilities of a Data Engineer is to monitor and optimize systems and data flows for **performance** and **availability**.

Data Pipelines

A data pipeline typically runs with a combination of complex tools and can face several different types of performance threats.

Performance threats include:

- Scalability in the face of increasing data sets and workloads
- Application failures
- Scheduled jobs not functioning accurately
- Tool incompatibilities

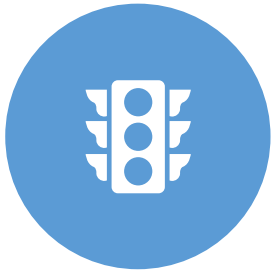


Data Pipelines - Performance Metrics

LATENCY - THE TIME IT TAKES FOR A SERVICE TO FULFILL A REQUEST

FAILURES - THE RATE AT WHICH A SERVICE FAILS

RESOURCE **UTILIZATION** AND UTILIZATION PATTERNS



TRAFFIC - NUMBER OF USER REQUESTS RECEIVED IN EACH PERIOD

Data Pipelines - Troubleshooting

Collect

Collect information about the incident to ascertain if the observed behavior is an issue.

Check

Check if you're working with all the right versions of software and source

Check

Check the logs and metrics early on in your troubleshooting process to isolate whether an issue is related to infrastructure, data, software, or a combination of these.

Reproduce

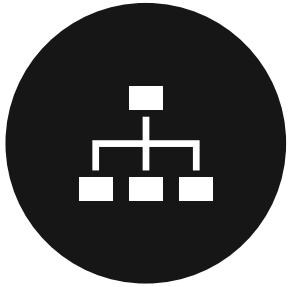
Reproduce the issue in a test environment. This can be an iterative and time-consuming process.

Database Optimization for Performance

Performance Metrics for Databases:

- System outages
- Capacity utilization
- Application slowdown
- Performance of queries
- Conflicting activities and queries being executed simultaneously
- Batch activities causing resource constraints

Database Optimization for Performance



Capacity Planning

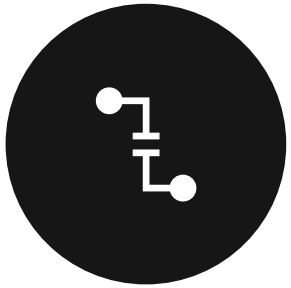
Determining the optimal hardware and software resources required for performance.

Database Indexing

Locating data without searching each row in a database resulting in faster querying.

Database Partitioning

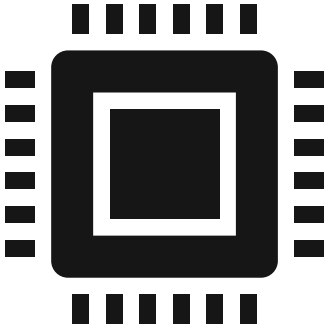
Dividing large tables into smaller, individual tables, improving performance and data manageability.



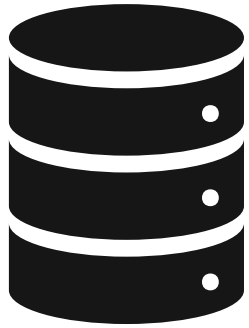
Database Normalization

Reducing inconsistencies arising out of data redundancy and anomalies arising out of update, delete, and insert operations on databases.

Monitoring Systems



Monitoring and alerting systems help us collect quantitative data about our systems and applications in real time.



These systems give visibility into the performance of data pipelines, data platforms, databases, applications, tools, queries, scheduled jobs, and more.

Monitoring Systems

Database Monitoring Toots take frequent snapshots of the performance indicators of a database.

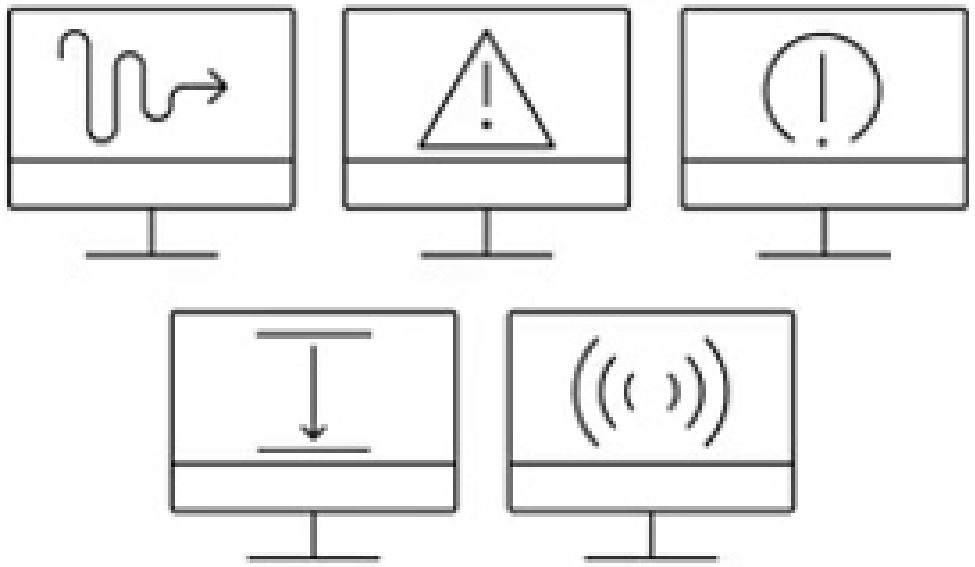
- This helps to:
  - Track when and how a problem started to occur.
  - Isolate and get to the root of the issue.
- **Application Performance Management Tools** measure and monitor the performance of applications and number of resources utilized by each process. This helps in proactive allocation of resources to improve application performance.
- **Tools for Monitoring Query Performance** gather statistics about query throughput, execution, performance, resource utilization and utilization patterns for better planning and allocation of resources.
- **Job-level Runtime Monitoring** breaks up a job into a series of logical steps which are monitored for completion and time to completion.
- **Monitoring Amount of Data being Processed** through a data pipeline helps to assess if size of workload is slowing down the system.

Maintenance Schedules

Preventive maintenance routines generate data that we can use to identify systems and procedures responsible for faults and low availability.

These routines can be:

- Time-based - Planned as scheduled activities at pre-fixed time intervals.
- Condition-based - Performed when there is a specific issue or a decrease in performance.




Thank
you