DevOps – Lecture 11

DataOps

24 Nov 2021

Chinmaya Dehury

Chinmaya.Dehury@ut.ee
OUTLINE

- Data-intensive application
- Data flows in data-intensive application
- Data-intensive application management, Challenges, participants
- What is DataOps?
- Benefits of DataOps
- DataOps participants
  - Data Analyst
  - Data Engineer
  - Data Scientist
- DataOps Dimensions
  - Total Quality Management (TQM)
  - Lean manufacturing
- DataOps Principles
- Challenges of Data Analytics
- Key steps in DataOps Implementations
- DevOps and DataOps
- Tools in DataOps
We call an application data-intensive if the data is its primary challenge - the quantity of data, the complexity of data, or the speed at which it is changing.
Data-intensive application

- Application involved in
  - Processing large volume of data
  - I/O operations
- Movement and manipulation of data.
- Data?
  - Text, Audio, Video...
  - Variety of data: Structure, Unstructured, semi-structures
  - Velocity of data: Batch data, stream data
- Data-intensive application mostly focuses on **Big data**
Data-intensive application

- Data-intensive application mostly focuses on **Big data**
- Some disciplines
  - Astronomy,
  - atmospheric science,
  - medicine,
  - genomics,
  - biologic,
  - Biogeochemistry
  - intelligent transportation systems
  - social network analysis
  - recommender systems
  - ..... and many more
How data flows in data-intensive application?

The Data-pipeline

Data Sources
- RDBMS
- Social media
- Sensors/IoT
- Machines
- ERP
- Mobile devices

Data Capture
- Batch ingest
- Stream ingest

Data Storage
- Data warehouse
- Data Lake

Data Process/transform
- Batch processing
- Stream data
- ML tools
- Task Orchestration

Result Storage
- Feature
- Data catalog
- Real-time store

Data Consumer
- Report
- ML APPs
- IoT APPs
- Analytics
- BI
- Alerts
- Services

Data Product
What are different Operations?

So what are the possible operations in this pipeline?

- How you collect the data from different sources?
  - e.g. Crawling tweets, facebook posts, review data, from different websites
  - E.g. Getting data from thousands of Sensors
- Storage
  - external/internal storage service provider
- Preparing the data from analysis
- Data analytics job
- Designing the models
- Preparing the result for data consumer
- How to setup the right tool set?
- And many more....
Data-intensive application management

Hard to find the root cause for:
- Application failure
- Data pipeline missing SLAs
- Huge Cloud cost
- Problem can be on Code or on the cluster...etc.

Root cause could be somewhere:
- Bad Data merging
- Wrong container size
- Data layout and file format
- Machine degradation
- Scheduling setting
- Wrong network setting
- Configuration setting
- ...

LTAT.06.015 : DevOps - Lec 11 - DataOps
Data-intensive application Challenges

Challenges
• Data demand in every fields
• Very high complexity of data pipeline
• Lack of experts
• Data team

Business outcomes due to above challenges:
• Slower delivery
• Too many defects
• High cost
• Poor customer satisfaction
Who are the participants?

So what are the possible operations in this pipeline?

• How you collect the data from different sources?
  - e.g. Crawling twitter, facebook post, review data, from different websites
  - Getting data from thousands of Sensors

• Storage
  - external/internal storage service provider

• Preparing the data from analysis
• Data analytics job
• Designing the models
• Preparing the result for data consumer
• How to setup the right tool set?
• And many more....

Possible list of Participants

• Data engineer
• Data scientists
• Data analyst
• Architects
• Developer

• Production team
• Monitoring
• Consumer
Can you deliver the product or extract the value out of data very quickly, with

- Such large pipeline
- So many potential complex problems
- So many stakeholders
DataOps
What is DataOps?

DataOps ➔ Data + Operations

“DataOps is a state of mind”

“It is about continuous improvement”

- Wayne Eckerson, President@Eckerson Group

“Data is a Collaborative data management practice focused on improving the communication, integration and automations of data flows between data manager and consumers across.”

- Gartner

“DataOps seeks to reduce the end-to-end cycle time of data analytics, from the origin of ideas to the literal creation of charts, graphs and models that create value.”

- DataOps Manifesto
What is DataOps?

“*This uses technology to automate data delivery with appropriate level of security, quality and metadata to improve the use and value of data in a dynamic environment.*”

- Gartner

**Goal:** to create analytics in the individual development environment, advance into production, receive feedback from users and then continuously improve through further iterations

- Data manifesto
Benefits of DataOps

• End-to-end efficiency and observability
• Real-time collaboration
• Faster product delivery
• Fewer data defects
• Greater code reuse

Business outcomes:
• Slower delivery  Faster
• Too many defects  lesser now
• High cost  cheaper
• Poor customer satisfaction  happy customer
Who are the DataOps participants?

So what are the possible operations in this pipeline?

• How you collect the data from different sources?
  • e.g. Crawling twitter, facebook post, review data, from different websites
  • Getting data from thousands of Sensors
• Storage
  • external/internal storage service provider
• Preparing the data from analysis
• Data analytics job
• Designing the models
• Preparing the result for data consumer
• How to setup the right tool set?
• And many more....

Possible list of Participants

Data Team
• Data engineer
• Data scientists
• Data analyst
• Architects
• Developer

Ops Team
• Production team
• Monitoring
• Consumer
Who are the DataOps participants?

**Data Analyst**
- typically gathers data to identify trends
- performing statistical analyses to help answer questions and solve problems
- make queries to relational databases
- may also clean data
- put data in a usable format
- discarding irrelevant or unusable information
- Handle missing data
- spend more time on routine analysis
- generates reports regularly
Who are the DataOps participants?

**Data Engineer**
- prepare data
- develop, constructs, tests & maintain complete architecture
- ability to create and integrate APIs
- Knowledge on data pipelining and performance optimization
- Deploy ML & statistical models
Who are the DataOps participants?

Data Scientist

- Involved with designing data modeling processes
- Focused on developing new tools and methods to extract the information
- Create algorithms
- Create predictive models
- May design the way data is stored, manipulated and analyzed
DataOps Dimensions

**Agile**
- Self-organizing teams
- Business engagement
- Short sprints
- Regular retrospective

**Total Quality Management**
- Continuous measurement
- Cont. monitoring
- Benchmarking
- Cont. improvement

**DevOps**
- Team-based development
- Version control
- CI/CD
- Test-drive development

**Lean manufacturing**
- Focus on customer
- Waste and inefficiency identification
- Simplify and standardize process
- Automation
Total Quality Management (TQM)

• An ongoing process of detecting and reducing or eliminating errors.

• Used to streamline supply chain management, improve customer service, and ensure that employees are trained.

• **Total employee involvement:** Aims to hold all parties involved in the production process accountable for the overall quality of the final product or service.
Lean manufacturing

• In the late 1940s, **Toyota** put the foundations of Lean manufacturing.

• **Aimed** to reduce processes that don’t bring value to the end product.

• Significant improvements in **productivity, efficiency, cycle time, and cost-efficiency**

• E.g.: Restaurants prepare the dishes based on only customers’ need

https://kanbanize.com/lean-management/what-is-lean-management
DataOps Principles

1. Continually Satisfy Your Customer
2. Value Working Analytics
3. Embrace Change
4. It’s A Team Sport
5. Daily Interactions
6. Self-organize
7. Reduce Heroism
8. Reflect
9. analytics Is Code
10. Orchestrate
11. Make It Reproducible
12. Disposable Environments

DataOps Principles

1. Continually Satisfy Your Customer
2. Value Working Analytics
3. Embrace Change
4. It’s A Team Sport
5. Daily Interactions
6. Self-organize
7. Reduce Heroism
8. Reflect
9. analytics Is Code
10. Orchestrate
11. Make It Reproducible
12. Disposable Environments
13. Simplicity
14. Analytics Is Manufacturing
15. Quality Is Paramount
16. Monitor Quality And Performance
17. Reuse
18. Improve Cycle Times

Challenges of Data Analytics

1. Always changing Goal
   • Users/clients do not know all possible insights.
   • Never ending questions from the client.
   • Clients need everything in short time in this competitive market.

2. Data Lives in Silos
   • E.g. data related to orders, deliveries, returns, website page views, mobile app navigations, downloads, clicks, metrics, audio logs, social media and more
   • Different databases
   • Different platforms
   • Different APIs

3. Data formats are not Optimized
   • E.g. Some timestamp in millisecond and some in nanosecond precision
   • Recall practice session, where some student changed the data to convert the epoch time to human readable format.
Challenges of Data Analytics

4. Data Errors
   • It is hard to get data without any error either from internal or external sources
   • E.g. duplicate data, missing data, outlier data points.

5. Bad data
   • This may bring bad impression
   • Client may now blame the data-analytics team not the data itself

6. Never-ending maintenance of Data pipeline
   • New data source
   • Schema enhancement
   • New analytic job...

7. Manual Process Fatigue
   • Difficult to have end-to-end automated pipeline
   • Intermediate manual process may burnout the team
Key steps in DataOps Implementations

1. Add Data And Logic Tests
   • e.g. number of row after join or cross product operation
   • e.g. number of failed transaction
   • Continuously monitor for errors and anomalies

2. Use Of VCS
   • related tools e.g. data prep tool, reporting tool, visualization tools should be in VCS
   • scripts, source code, algorithms, html, configuration files, parameter files, containers and other files
   • heavy use of branching and merging

3. Use Of Multiple Environments
   • Local development environment
   • use of subset of data incase local environment doesn't have enough resource

4. Reuse & Containerize
   • Not only the code, but also the intermediate result should be shared among other members
   • Containerize the code as well

5. Parameterize Your Processing
   • Make the analytic algorithm flexible enough to take the variables in runtime
DevOps and DataOps

• Optimizing code builds and delivery is a part of data analytics’ job.
• DataOps introduces Agile Development
  • Data Team, IT team and the client work together efficiently and effectively
• DataOps borrows the methodology from DevOps.
  • Also from Agile development and Lean manufacturing
• Stakeholder
  • In DataOps, the main stakeholder is the data team (Data analyst, data engineer or data scientist)
  • One of the primary focus is on building models and visualization
DevOps and DataOps

Development and Deployment Processes

DevOps Process
- Develop
- Build
- Test
- Deploy
- Run

DataOps Process
- Sandbox Management
- Develop
- Orchestrate
- Test
- Deploy
- Orchestrate
- Monitor

Src: https://dataopsmanifesto.org/en/
DevOps and DataOps

Development and Deployment Processes

This includes several tools, languages and vendors for data engineering, data science, BI, visualization and governance.

Access -> Transform -> Model -> Visualization -> Report

Here it is referring to automated testing, Monitoring and lean manufacturing

Src: https://dataopsmanifesto.org/en/
DevOps and DataOps

Complexity in DataOps

1. Sandbox Management
   • An isolated environment
   • Specially for new feature
   • This is the developer’s environment with several tools, libraries

2. Test data management
DevOps and DataOps

Complexity in DataOps

1. Sandbox Management
2. Test data management
   - Copy of original data
   - May copy a subset of original data
   - Comes with security and licensing restrictions.

https://medium.com/data-ops/dataops-is-not-just-devops-for-data-6e03083157b7
# Tools in DataOps

<table>
<thead>
<tr>
<th>Category</th>
<th>Tools</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collaboration</td>
<td>slack, Jira, Trello, Microsoft Teams</td>
</tr>
<tr>
<td>Deployment</td>
<td>kubernetes, Jenkins, Azure DevOps, Bash</td>
</tr>
<tr>
<td>Orchestration</td>
<td>Azure Automation, puppet, Ansible, Airflow</td>
</tr>
<tr>
<td>Testing and Monitoring</td>
<td>Datadog, Azure Application Insights, Azure Log Analytics</td>
</tr>
</tbody>
</table>
## Tools in DataOps

<table>
<thead>
<tr>
<th>Category</th>
<th>Tools</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Capture</td>
<td>SQL Server, Kafka, Azure Stream Analytics</td>
</tr>
<tr>
<td>Data Storage</td>
<td>S3, Azure Data Lake, SQL, Google BigQuery, Azure Cosmos DB</td>
</tr>
<tr>
<td>Data Integration</td>
<td>Airflow, SQL Server, Azure Data Factory</td>
</tr>
<tr>
<td>Data Governance</td>
<td>Apache Atlas, Informatica, Collibra</td>
</tr>
<tr>
<td>Data Analytics</td>
<td>Power BI, Databricks, Azure ML</td>
</tr>
</tbody>
</table>
Lab Sessions

Dataset:
• Particulate Matter (PM) dataset for Delhi, India
• Area spanning 559 square kms.
• PM data recorded over three months from November 2020 to January 2021
• Data Source: https://www.cse.iitd.ac.in/pollutiondata/
• Data Attributes:
  • uid
  • dateTime
  • deviceId
  • lat
  • long
  • pm1_0
  • pm2_5
  • pm10
Lab Sessions

Use case to implement in Practice session (This is a draft version)

InfluxDB:
Data Source 1:
Location 1:
Container: TA’s VMs

InfluxDB:
Data Source 2:
Location 2:
Container: TA’s VMs

Data Preprocess:
Combine and regroup according to Device Id:
Container:

Data analytic 1:
Calculate hourly, daily, and weekly avg. + Push data to the different database:
Container:

Data analytic 2:
Calculate Air quality Index:
Container:

Local Intermediate storage:
InfluxDB:
Container: Student’s VM:

Grafana:
Container: Student’s VM:
Lab Sessions

Use case to implement in Practice session (This is a draft version)

- **InfluxDB**: Data Source 1: Location 1: Container: TA's VMs
- **InfluxDB**: Data Source 2: Location 2: Container: TA's VMs

**Data Preprocess**: Combine and regroup according to Device Id: Container:

**Data analytic 1**: Calculate hourly, daily, and weekly avg. + Push data to different database: container:

**Data analytic 2**: Calculate Air quality Index: Container:

**Data analytic 3**: Students' homework Find average in morning, noon, evening, and night: Container:

**Local Intermediate storage**: InfluxDB: Container: Student's VM:

**Grafana**: Container: Student's VM:
References


• https://dkproduction.staging.wpengine.com/blog/


• https://www.cse.iitd.ac.in/pollutiondata/

• https://www.dataops.rocks/

• https://dataopsmanifesto.org/en/

• https://docs.microsoft.com/en-us/azure/architecture/example-scenario/data-warehouse/dataops-mdw

• https://datakitchen.io/what-is-dataops/

• https://asq.org/quality-resources/total-quality-management

• https://kanbanize.com/lean-management/what-is-lean-management


Any Question?

THANK YOU