Data acquisition, migration and flow management

13 April 2021

Chinmaya Dehury

chinmaya.dehury@ut.ee
Outlines

• Data Acquisition
• Data Migration
• Data Pipeline solutions
• AWS data pipeline
• Apache Nifi
Data Acquisition

• Process of gathering, filtering, and cleaning data
• Data can be:
  • Text
  • Audio
  • Video
• 5V’s of the data:
  • Volume (size of the data)
  • Velocity (how fast the data is generated?)
  • Variety (Structured, Semi-Structured, Unstructured data)
  • Veracity (messy, quality, and accuracy?)
  • Value
Data Migration

• transferring data from one computer storage system to another. e.g. :
  • Transferring images from your smart phone to your laptop
  • Transferring data from old laptop to new one
  • Transferring data from Google Drive to Dropbox

• process of selecting, preparing, extracting, transforming data and transferring

• Usually thousands of data sources are involved
• Generated data are of small size
• Higher frequency of data generation
Data Migration

Challenges & Risks

• Data Loss

• Compatibility issues:
  • Storage Compatibility
  • Application compatibility
  • Platform compatibility (such as form on-premise to cloud),
  • Cloud compatibility
Data Migration

2 Broad categories
• Online
• Offline

Factors to consider:
• Type of workload:
  • Databases, virtual machines (VMs), Backups, etc
• Amount of data
  • Imagine migrating some Petabytes of data online
  • Imagine migrating few GBs of data in offline mode
• Speed to completion:
  • For online migrations: amount of data
  • For offline migrations: shipping time
Then What is Data Pipeline?
Data Pipeline

Pipeline approach for computer instruction execution:

Pipeline approach in manufacturing:
Pipeline approach in logistic:
Pipeline approach for handling the data acquisition, migration and its flow.
Data Pipeline (DP)

Large data processing task

Data in → Data processing → Data out

Data in → Data out
Data Pipeline (DP)

• A system for moving data from one system to another.
• Encompasses ETL as a subsystem
• Transformation of data is optional
• May be processed in real-time or in batch manner
Data Pipeline properties

1. Low Event Latency
2. Scalability
3. Interactive Querying
4. Versioning
5. Monitoring
6. Testing
Types of data pipeline solutions

1. Batch
2. Real-time
3. Cloud native
4. Open source
1. Amazon Data pipeline

2. Apache Nifi
Data Pipeline Technologies

1. Amazon Data pipeline

2. Apache Nifi
Amazon Data Pipeline

• A web service for reliable process and movement of data
• Focus is on AWS compute and storage services
• AWS services such as Amazon S3, Amazon RDS, Amazon DynamoDB, and Amazon EMR
• Data processing workloads can be
  • fault tolerant
  • repeatable
  • highly available
Amazon Data Pipeline: An Example

Configuration
Default

Schedule
Every 1 Day

S3Datanode
IntermediateDocDir

CopyActivity
CopyActivity

S3Datanode
DataDirectory

EMRCluster
DefaultEMRCluster1

PigActivity
PigProcessDocuments

S3Datanode
OutputDirectory

Acquisition, migration and flow management
Amazon Data Pipeline: An Example

**Configuration**
- Pipeline Configuration
- Execution Schedule
  - Schedule: Every 1 Day

**Data Node**
- S3Datanode
  - IntermediateDocDir
  - CopyActivity
  - CopyActivity

**AWS Resource**
- EMRCluster
  - DefaultEMRCluster1

**Pipeline Activity**
- S3Datanode
  - DataDirectory

**Data Node**
- PigActivity
  - PigProcessDocuments

**Output Directory**
- S3Datanode
  - OutputDirectory
Amazon Data Pipeline

1. Major components
   I. DataNodes
   II. Activities

2. Additional components
   I. Schedules
   II. Preconditions
   III. Resources
1. Major components

I. **DataNodes**: It specifies the name, location, and format of the data sources such as Amazon S3, Dynamo DB, etc.
   i. DynamoDBDataNode
   ii. SqlDataNode
   iii. RedshiftDataNode
   iv. S3DataNode
   v. SqlDataNode

II. **Activities**: Activities are the actions that perform the SQL Queries on the databases, transforms the data from one data source to another data source.
Amazon Data Pipeline

1. Major components

I. DataNodes

II. Activities

i. CopyActivity
ii. EmrActivity
iii. HadoopActivity
iv. HiveActivity
v. HiveCopyActivity
vi. PigActivity
vii. RedshiftCopyActivity
viii. ShellCommandActivity
ix. SqlActivity
Amazon Data Pipeline

1. Major components
   I. DataNodes
   II. Activities

2. Additional components
   I. Schedules: Schedule defines the timing of a scheduled event, such as when an activity runs.
2. Additional components

I. Schedules

II. Preconditions: A condition that must be true before an activity can run. E.g., check if the data is present on the source before attempting to run CopyActivity.

A. System-managed Precondition:
   a) DynamoDBDataExists
   b) DynamoDBTableExists
   c) S3KeyExists, etc..

B. User-managed precondition
   a) Exists: Checks whether a data node exists.
   b) ShellCommandPrecondition: Unix/Linux shell command that can be run as a precondition
Amazon Data Pipeline

2. Additional components

I. Schedules

II. Preconditions

III. Resources: refer to the computational resource that performs the work that a pipeline activity specified

I. **Ec2Resource**: An EC2 instance

II. **EmrCluster**: An Amazon EMR cluster
Data Pipeline Technologies

1. Amazon Data pipeline

2. Apache Nifi
Apache Nifi Data Pipeline

• Open-source, under the Apache Software Foundation
• Automates and manages the flow of data between systems
• Web-based User Interface for creating, monitoring, & controlling data flows.

• Clients [src]:
  • Micron: Semiconductor Manufacturing
  • Payoff: Financial Wellness (fintech)
  • Slovak: Telekom Telecommunications
  • Looker: SaaS & Analytics Software
  • Hastings Group: Insurance
  • and many more....

• Latest version 1.13.2 (as on April 2021)
Apache Nifi Data Pipeline

**Key Features**

**Flow Management:**
- Data Buffering
- Prioritized Queuing
- Guaranteed Delivery

**Ease of Use:**
- Flow Templates
- Data Provenance
- Fine-grained history
Apache Nifi Data Pipeline

Key Features

Security
• System to System
• User to System
• Multi-tenant Authorization

Extensible Architecture
• Extension
• Site-to-Site Communication Protocol
Apache Nifi Data Pipeline

1. Major components
   I. Processors (execute the task)
   II. Queue (between processors)

2. Additional components
   I. Input Port
   II. Output Port
   III. Process Group (Groupism of multiple components such as processors)
   IV. Remote Process Group
   V. Template
Apache Nifi Data Pipeline

Key concepts

1. Process Group
2. Flow
3. Processor
4. Flowfile (represents a single piece of data)
5. Event
6. Data provenance
Apache Nifi Data Pipeline

NiFi Architecture

Local Storage

Source: https://www.tutorialspoint.com/apache_nifi/apache_nifi_basic_concepts.htm
stores the metadata of the FlowFiles during the active flow.

holds the actual content of the FlowFiles.

what happened to a particular data object (FlowFile) is kept in here.

Src: https://www.tutorialspoint.com/apache_nifi/apache_nifi_basic_concepts.htm
1. Major components
   I. Processors

   283 processors

   **Apache Nifi - Processors**

   ![Add Processor](image)

   **Major components**

   I. Processors

   283 processors
1. Major components
   I. Processors

Different States of a Processor:
Start, Stop, Enable, & Disable

Disable processor can not be started.
When a group of Processors is started, this (disabled) Processor should be excluded
Apache Nifi – Processors Setting

1. Major components
   I. Processors

Configuring a Processor

**SETTING:**

**Penalty duration:** Time to wait, when the data can not be processed for some reason.

**Yield Duration:** Time to wait, when the process can not progress.

**Bulletin level:** Level of bulletin, Nifi will display in the user interface.
Apache Nifi – Processors Scheduling

1. Major components
   I. Processors

Configuring a Processor

Scheduling:

Time vs Event vs CRON Driven

Concurrent Tasks: Number of FlowFiles should be processed by this Processor at the same time.
Apache Nifi – Processors Properties

1. Major components

   I. Processors

Configuring a Processor

Properties:

- Provides a mechanism to configure Processor-specific behavior.
- There are no default properties.
Apache Nifi – Processor categories

Different categories of processors

• **Data Ingestion Processors:** GetFile, GetHTTP, GetFTP, etc.

• **Routing and Mediation Processors:** RouteOnAttribute, RouteOnContent, ControlRate, RouteText, etc.

• **Database Access Processors:** ExecuteSQL, PutSQL, PutDatabaseRecord, ListDatabaseTables, etc.

• **Attribute Extraction Processors:** UpdateAttribute, EvaluateJSONPath, ExtractText, AttributesToJSON, etc.

• **System Interaction Processors:** ExecuteScript, ExecuteProcess, ExecuteGroovyScript, ExecuteStreamCommand, etc.
Different categories of processors

- **Data Transformation Processors**: ReplaceText, JoltTransformJSON, etc.
- **Sending Data Processors**: PutEmail, PutSFTP, PutFile, PutFTP, etc.
- **Splitting and Aggregation Processors**: SplitText, SplitJson, SplitXml, MergeContent, SplitContent, etc.
- **HTTP Processors**: InvokeHTTP, ListenHTTP, etc.
- **AWS Processors**: GetSQS, PutSNS, PutS3Object, FetchS3Object, etc.
1. Major components

II. Queue
Apache Nifi - Queue

1. Major components

II. Queue

- To handle the large amount of data inflow.
- Possible to see the content, ID, Filename, FileSize etc of a flowfile.
Apache Nifi - Flow Template

**Templates:**
- Can be thought of as a reusable sub-flow.
- Any properties that are identified as being Sensitive Properties (such as a password that is configured in a Processor) will not be added to the template.
Apache Nifi - Flow Template

Templates:

Upload Template

Add Template
Apache Nifi - Flow Template

Templates:

Upload Template
Apache Nifi – Data Provenance

- Snapshots of each FlowFile.
- Event type, FlowFile Lineage Graph,
- Provenance event Details
- In-depth discovery of the chain of events.
Apache Nifi – Data Provenance

List of Events

Event type: RECEIVE, SEND, DROP, JOIN, CONTENT_MODIFIED, ATTRIBUTES_MODIFIED, FORK, CLONE, ROUTE, etc.

FlowFile Lineage Graph

Provenance event details
Provenance event Details

<table>
<thead>
<tr>
<th>DETAILS</th>
<th>ATributes</th>
<th>CONTENT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td>04/12/2021 13:46:33.595 UTC</td>
<td></td>
</tr>
<tr>
<td>Event Duration</td>
<td>&lt;1ms</td>
<td></td>
</tr>
<tr>
<td>Lineage Duration</td>
<td>00:00:00.014</td>
<td></td>
</tr>
<tr>
<td>Type</td>
<td>DROP</td>
<td></td>
</tr>
<tr>
<td>FlowFile Uuid</td>
<td>a6377ad4-7305-4bd6-8c92-7673c15b537b</td>
<td></td>
</tr>
<tr>
<td>File Size</td>
<td>100 bytes</td>
<td></td>
</tr>
<tr>
<td>Component Id</td>
<td>c638242b-017b-1000-cd4b-2ba25003731f</td>
<td></td>
</tr>
<tr>
<td>Component Name</td>
<td>PutFile</td>
<td></td>
</tr>
<tr>
<td>Component Type</td>
<td>Unknown</td>
<td></td>
</tr>
</tbody>
</table>

**Parent FlowFiles (0)**
No parents

**Child FlowFiles (0)**
No children
Apache Nifi – Data Provenance

FlowFile Lineage Graph

Event
FlowFile
Event whose graph was selected (red color)

Timestamp of the event

04/12/2021 13:46:33.593 UTC
RADON’s Data pipeline

- We in RADON, focusing on developing the data pipeline platform for data intensive applications.
- For serverless applications
- TOSCA model for data pipeline
- Atop Apache NiFi, Amazon data pipeline.
What next ???
Let’s move to lab session...
References


Thank you