UNIVERSITY OF TARTU
INSTITUTE OF COMPUTER SCIENCE

Cloud Computing – Lecture 12

Data streams, data flow pipeline management

28 April 2020

Chinmaya Dehury

Satish Srirama

Mobile & Cloud Lab

TARTU ÜLIKOOL · UNIVERSITAS TARTUENSIS · 1632

Outlines

- Data streaming
- Data pipeline
- Amazon data pipeline
- Apache Nifi

Data streaming

- Continuous flow of data
- Usually thousands of data sources are involved
- Generated data are of small size
- Higher frequency of data generation

Stream data processing use cases

- **Anomaly Detection**
  - Detect problems in real time (cyber intrusions, financial fraud,
  - Continuously collect and analyse network traffic, transactions, user behaviour
- **Predictive maintenance**
  - Collect and process performance data from deployed devices
  - Forecast potential faults and service disruptions, predict maintenance cycles
- **Clickstream analytics**
  - Collect and analyse user clicks, routes and behaviour
  - Extract frequent patters to improve user engagement
  - Personalized recommendations

Stream data processing frameworks

- **Frameworks/extensions specifically designed for:**
  - Low latency data processing
  - Dynamically process incoming data streams
  - Aggregate data processed at different time periods
  - Push results to external systems as output streams
- **Two main approaches/models:**
  1. Micro Batch processing
  2. Real Time stream data processing

Stream processing models

- **Micro Batch processing**
  - Collect incoming data into a batch/buffer
  - Processing one batch at a time
  - High throughput, High latency
  - Spark Streaming
- **Real time processing**
  - Process each incoming message right away
  - Low latency, lower throughput
  - Apache Storm, Apache Flink

Then What is Data Pipeline ?

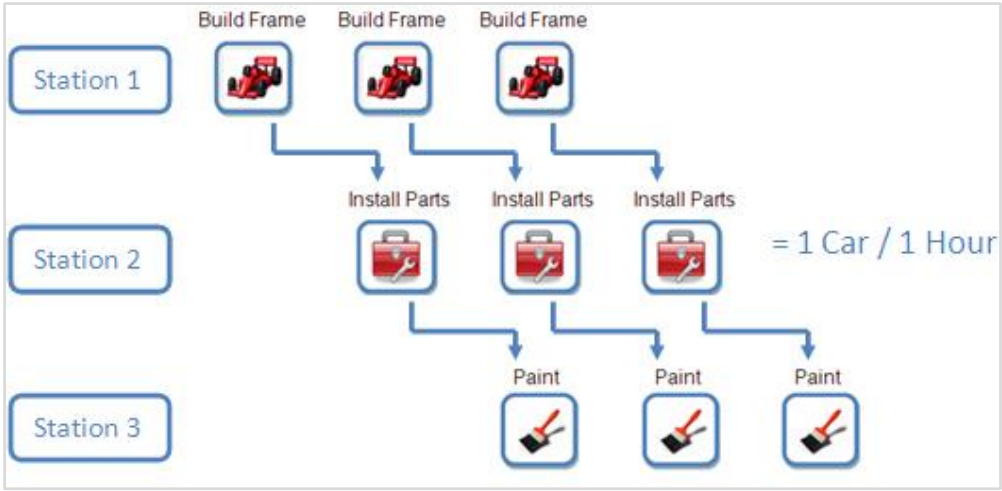
Data Pipeline

Pipeline approach for computer instruction execution:

	Clock cycle 1	2	3	4	5	6	7	8	9
lw $t0, 4($sp)	IF	ID	EX	MEM	WB				
sub $v0, $a0, $a1		IF	ID	EX	MEM	WB			
and $t1, $t2, $t3			IF	ID	EX	MEM	WB		
or $s0, $s1, $s2				IF	ID	EX	MEM	WB	
add $sp, $sp, -4					IF	ID	EX	MEM	WB

https://slideplayer.com/slide/8207220/

Pipeline approach in manufacturing:



http://www.ni.com/cms/images/devzone/tut/final.JPG

Data Pipeline

Pipeline approach in logistic:

Logistics Information:

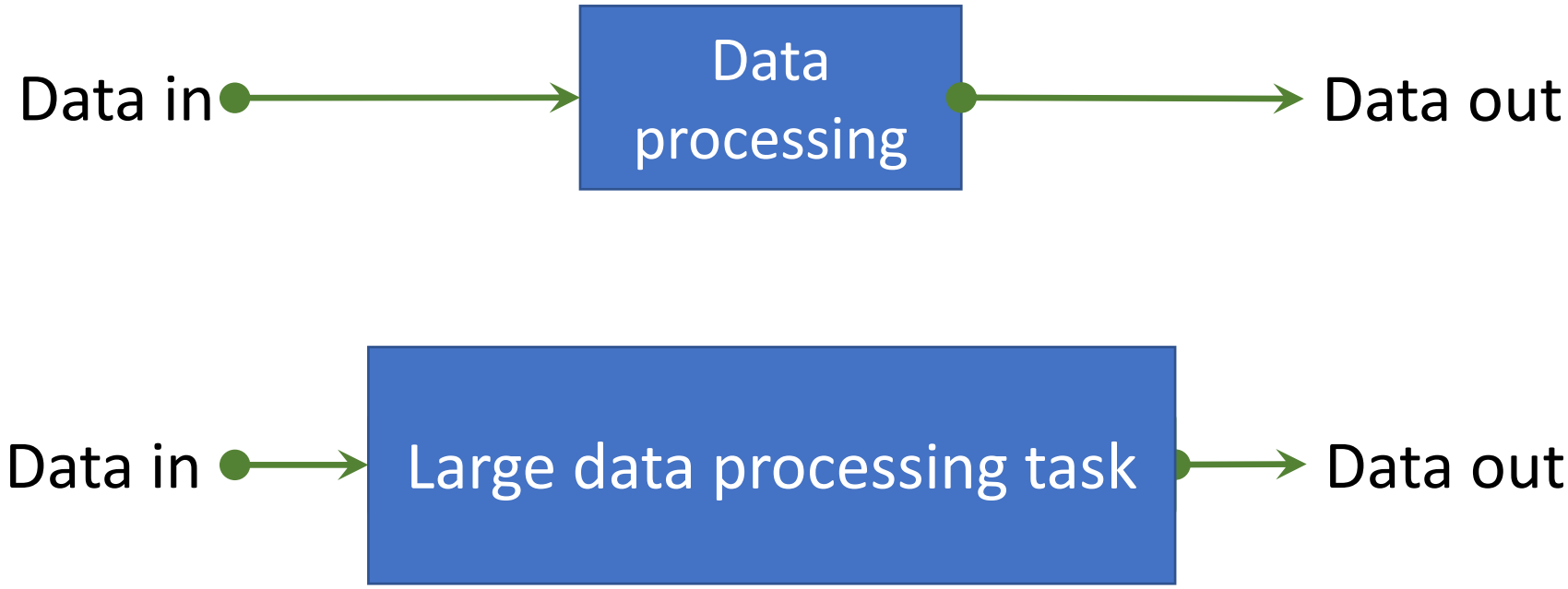
International Shipping Company	Tracking Number	Remarks	Details
菜鸟超级经济Global	S00000090969004		2019.11.26 19:37 (GMT-7): Departed country of origin 2019.11.26 14:37 (GMT-7): Shipment accepted by airline 2019.11.26 14:37 (GMT-7): Shipment left country of origin warehouse 2019.11.26 04:01 (GMT-7): Shipment at country of origin warehouse 2019.11.26 03:49 (GMT-7): Shipment dispatched Refresh Tracking information is available within 5-10 days. You can track your order here 菜鸟超级经济 Global.

View Delivery Detail

Data Pipeline

Pipeline approach for handling the flow and processing of data.

Data Pipeline (DP)



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

Data Pipeline Technologies

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

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

Amazon Data Pipeline

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

   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. HiveActivity
      iv. HiveCopyActivity
      v. PigActivity
      vi. RedshiftCopyActivity
      vii. ShellCommandActivity
      viii. 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.

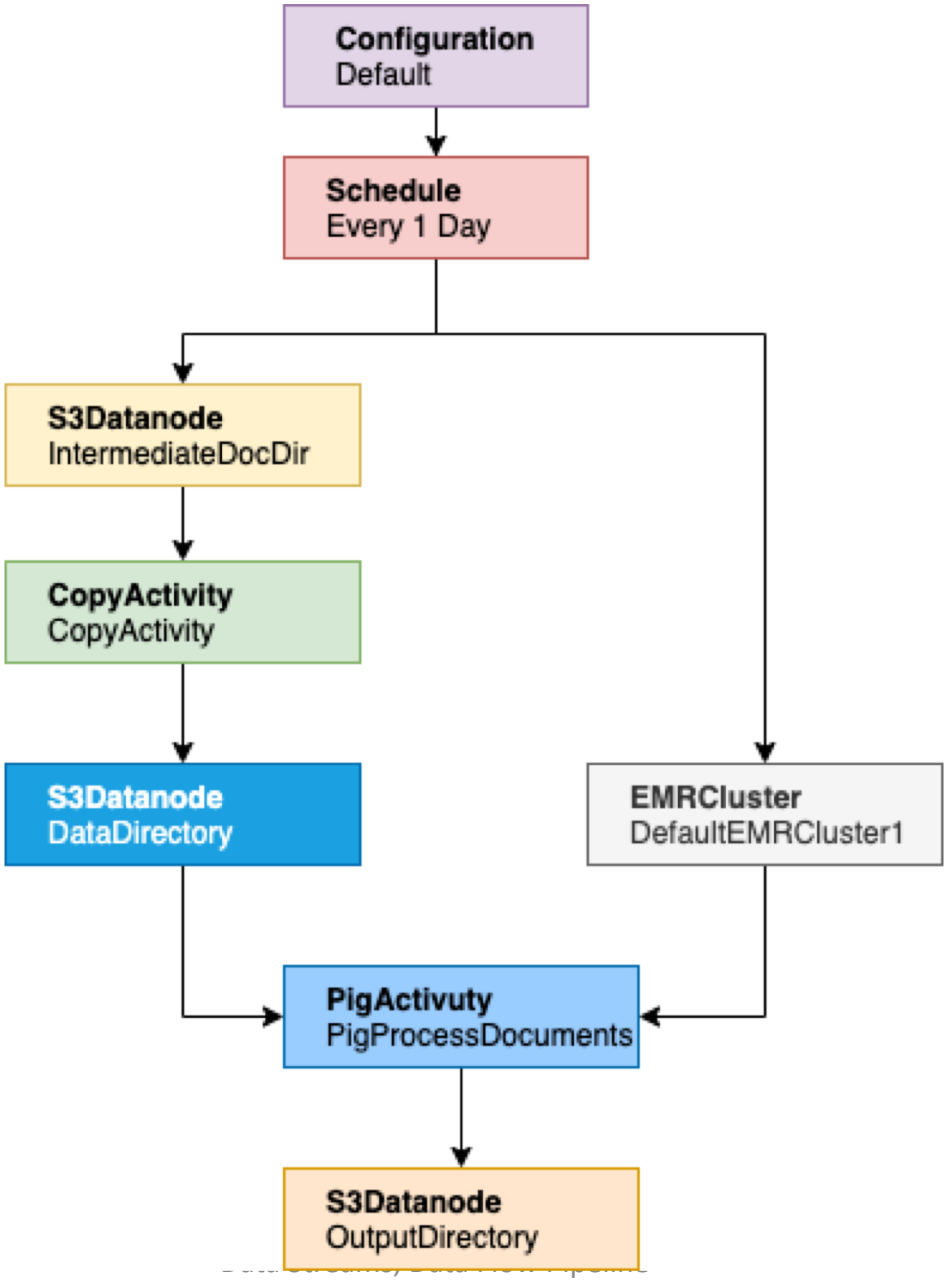
Amazon Data Pipeline

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

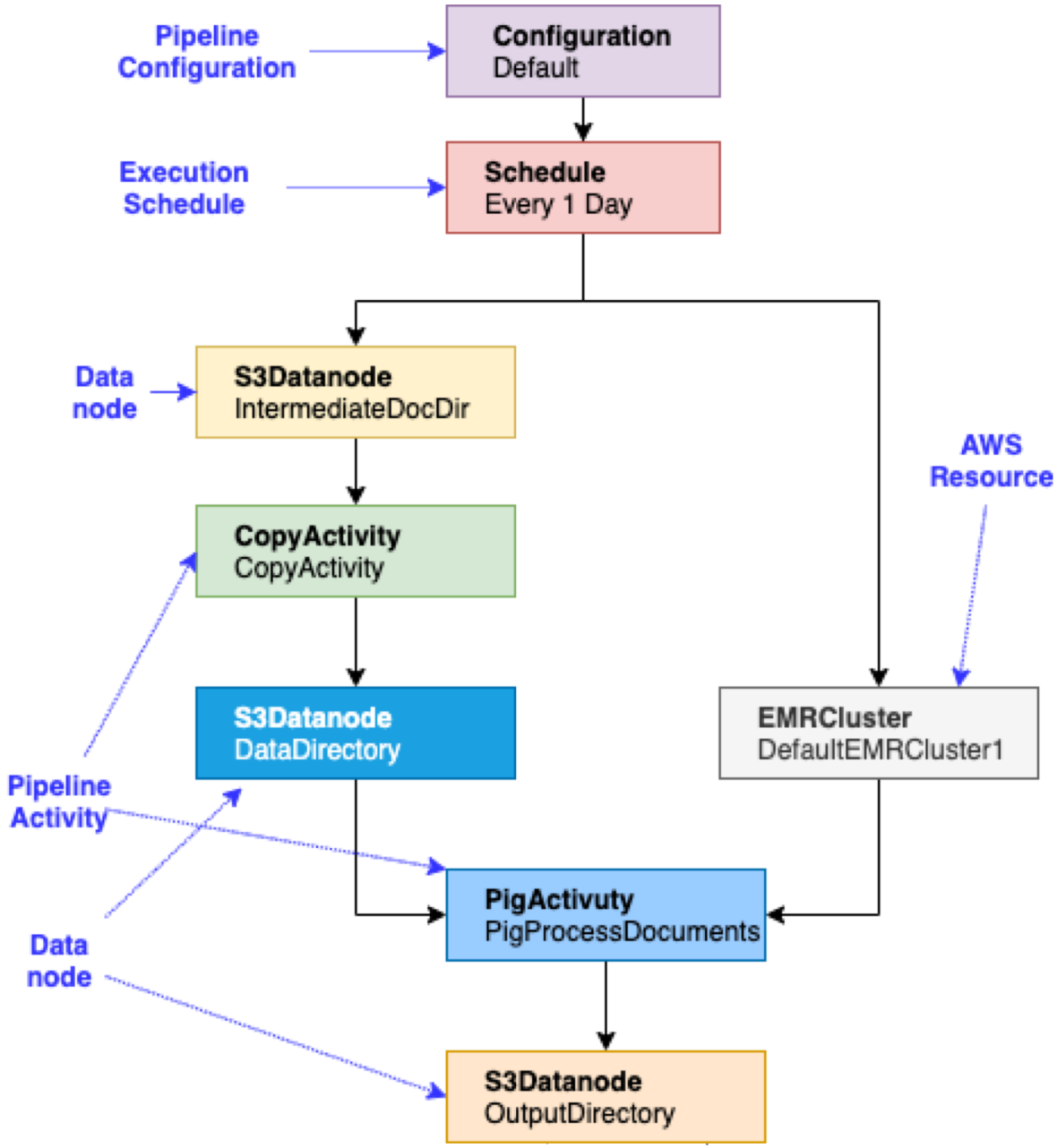
Amazon Data Pipeline

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

Amazon Data Pipeline : An Example



Amazon Data Pipeline : An Example



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 1.11.4

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
   II. Queue (between processors)
2. Additional components
   I. Input Port
   II. Output Port
   III. Process Group
   IV. Remote Process Group
   V. Template

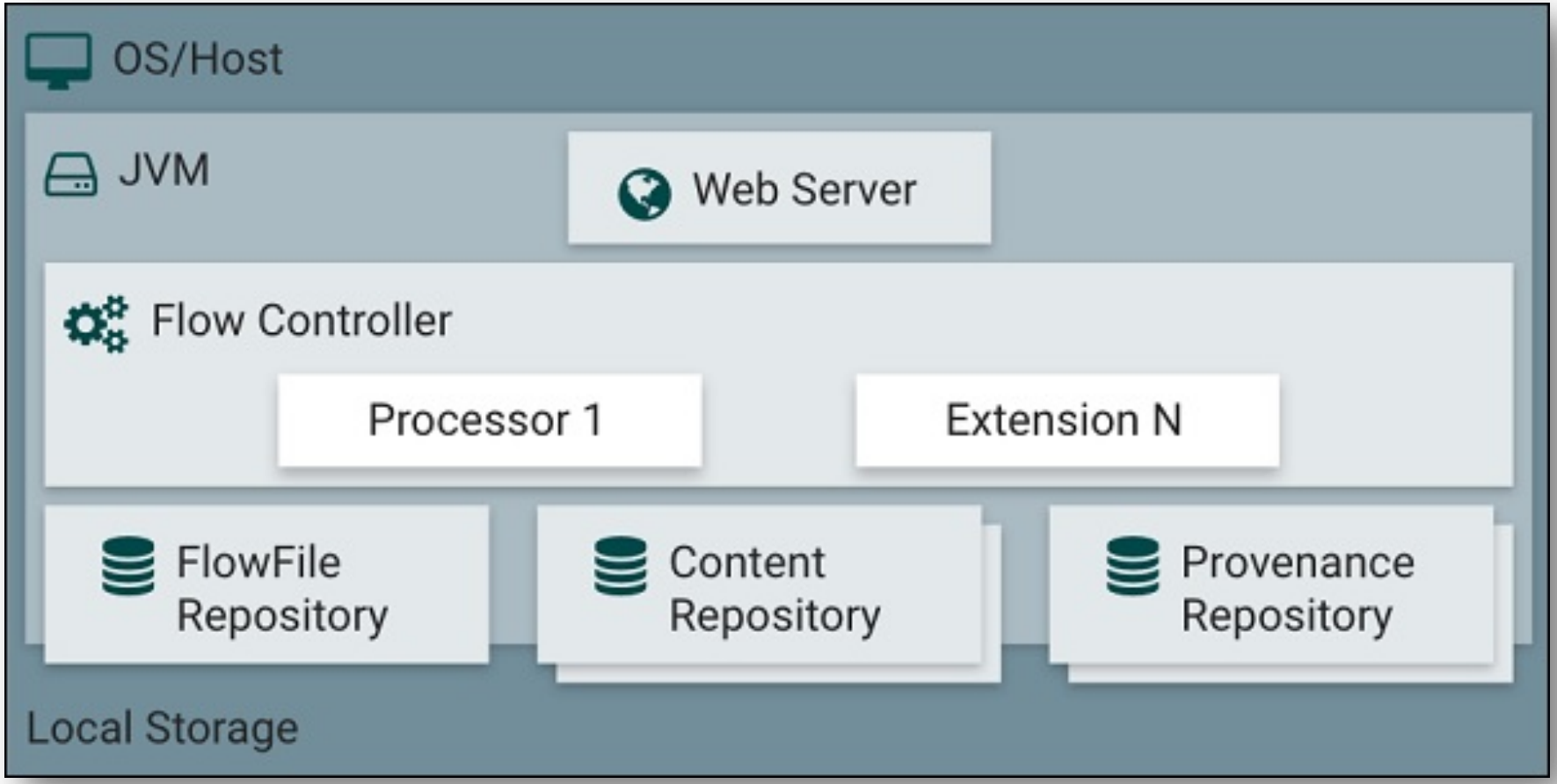
Apache Nifi Data Pipeline

Key concepts

1. Process Group
2. Flow
3. Processor
4. Flowfile
5. Event
6. Data provenance

Apache Nifi Data Pipeline

NiFi Architecture

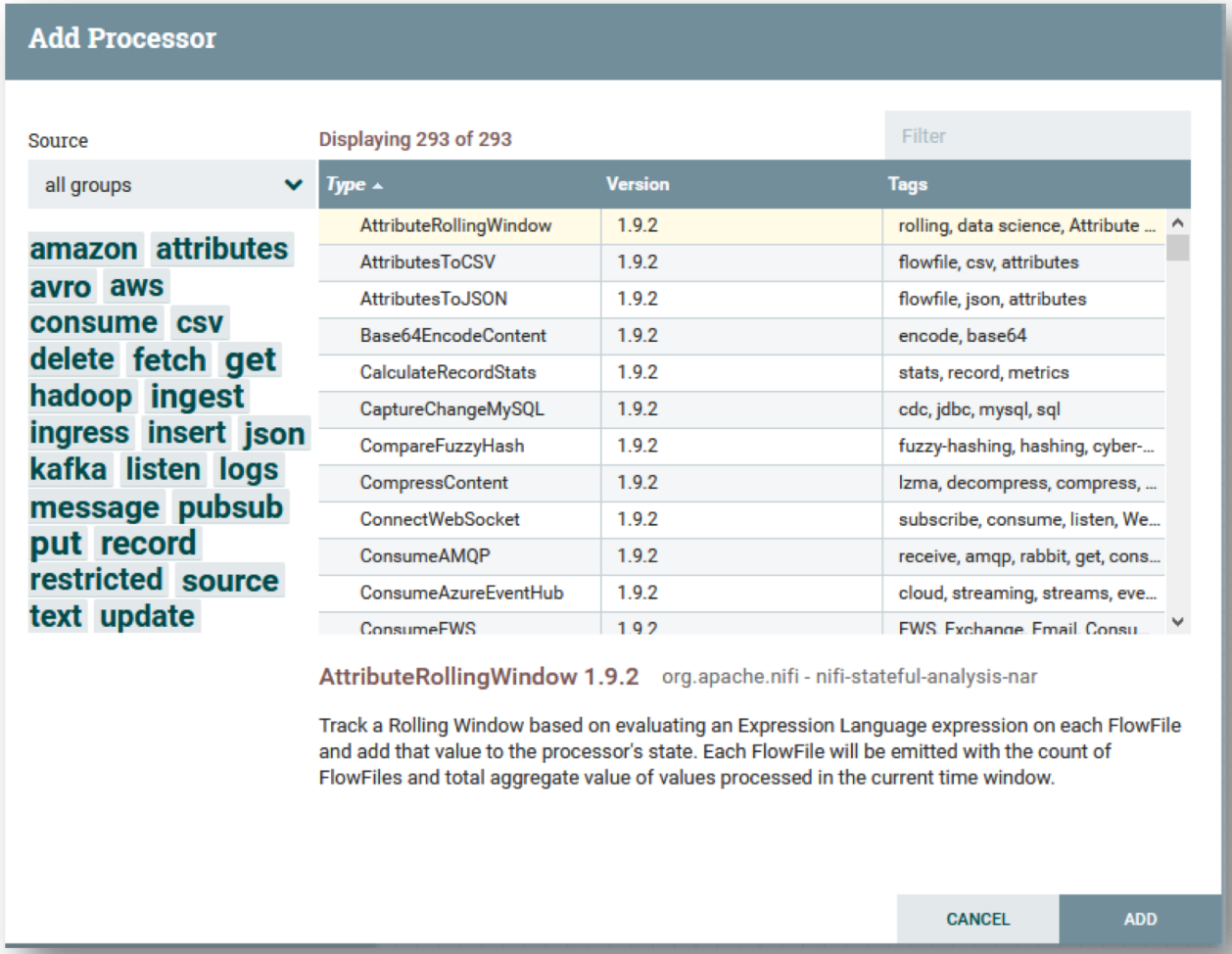


Src: https://www.tutorialspoint.com/apache_nifi/apache_nifi_basic_concepts.htm

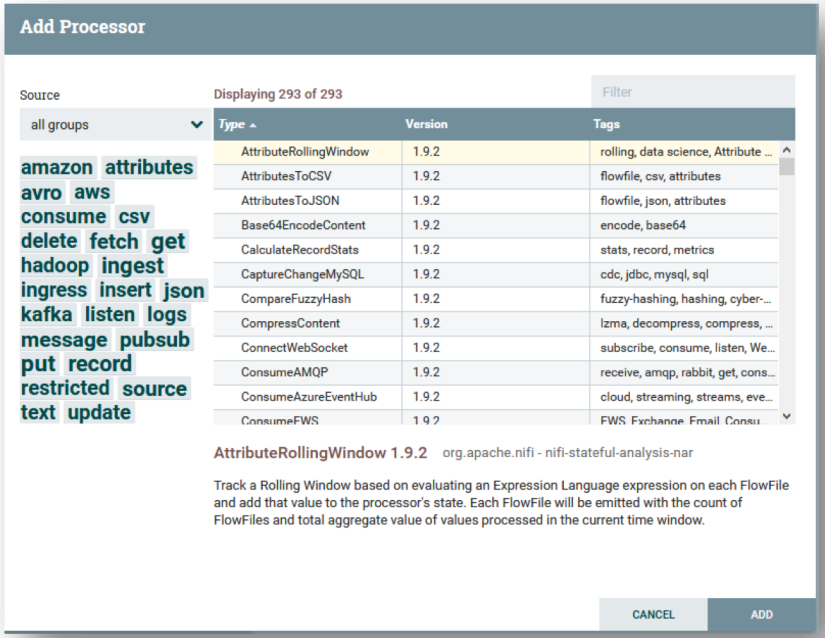
Apache Nifi Data Pipeline

1. Major components
   I. Processors

293 processors



Apache Nifi Data Pipeline



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 Data Pipeline

1. Major components

   I. Processors

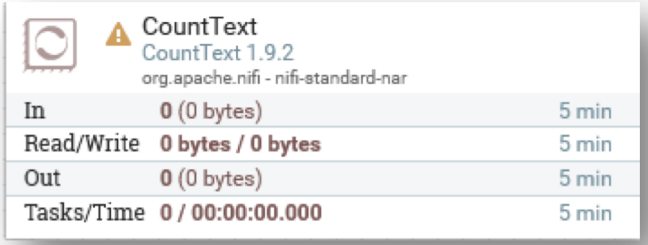
Configuring a Processor

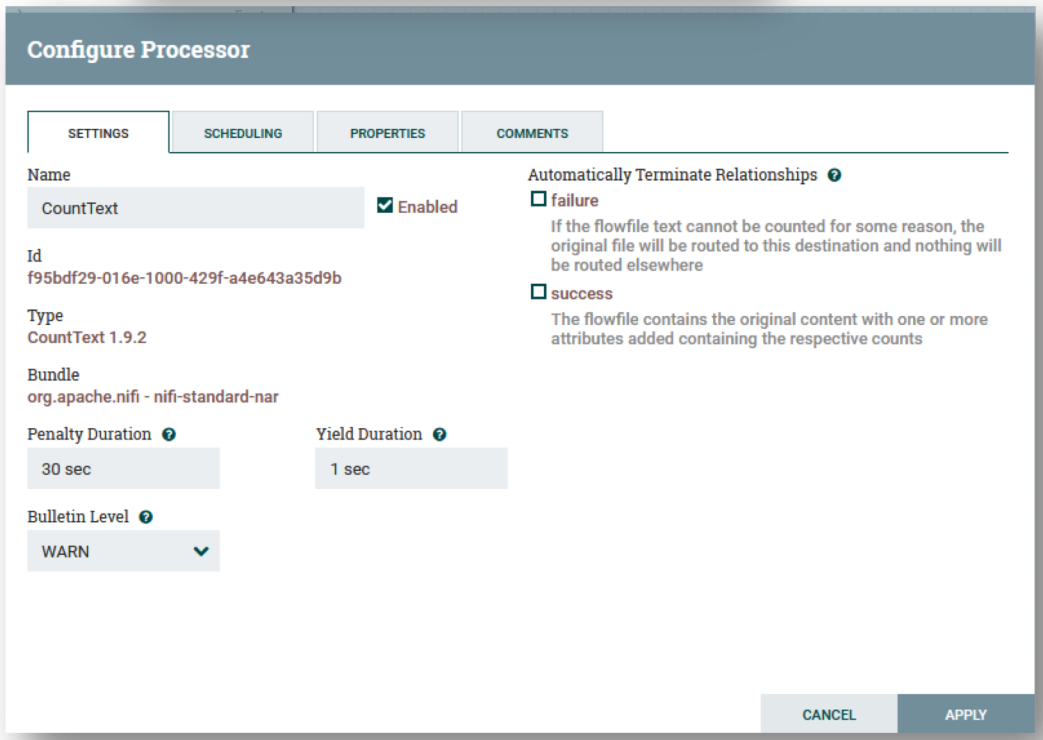
SETTING:

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

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

Bulletin level: Level of bulletin, Nifi will display in the user interface.





Apache Nifi Data Pipeline

1. Major components

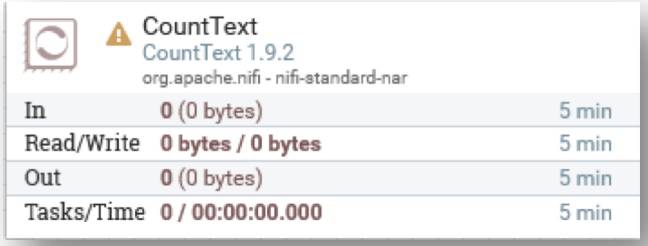
   I. Processors

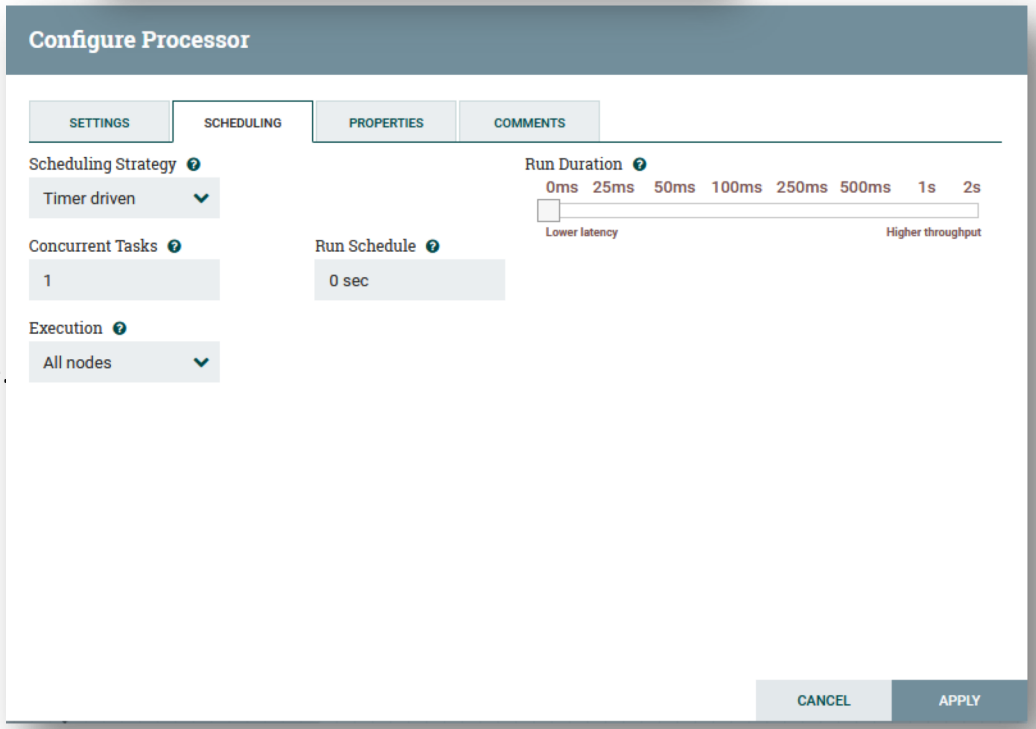
Configuring a Processor

Scheduling :

Time vs Event vs CRON Driven

Concurrent Tasks: Number of FlowFiles should be processed by this Processor at the same time.





Apache Nifi Data Pipeline

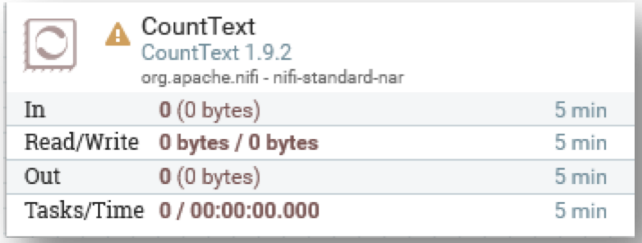
1. Major components

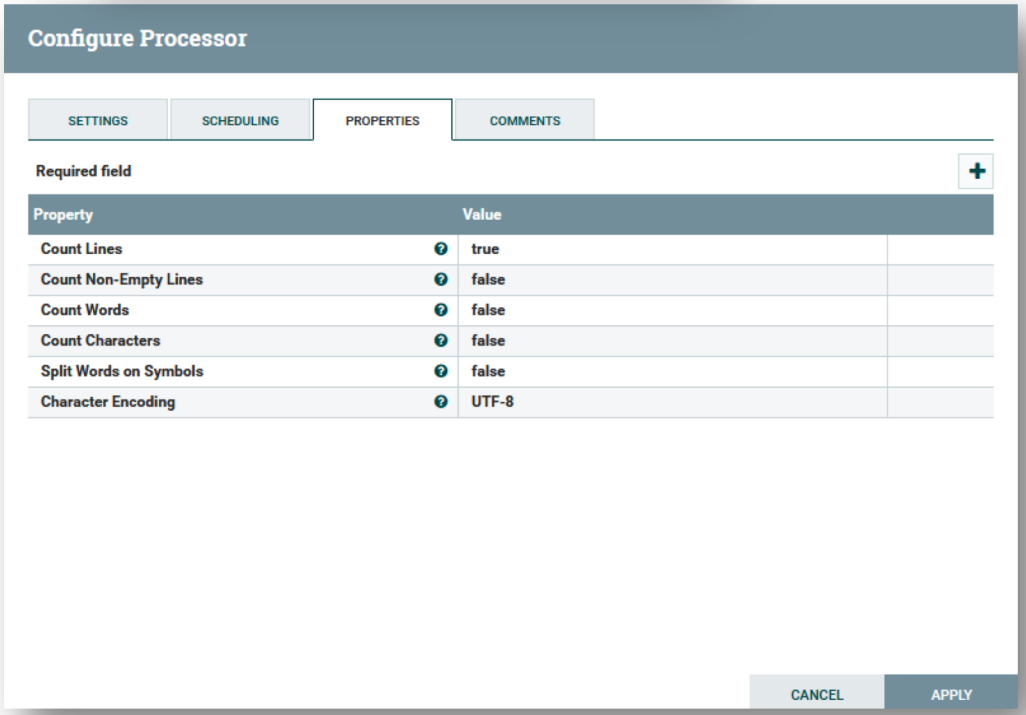
   I. Processors

Configuring a Processor

Properties :

- Provides a mechanism to configure Processor-specific behavior.
- There are no default properties.





Apache Nifi Data Pipeline

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

Apache Nifi Data Pipeline

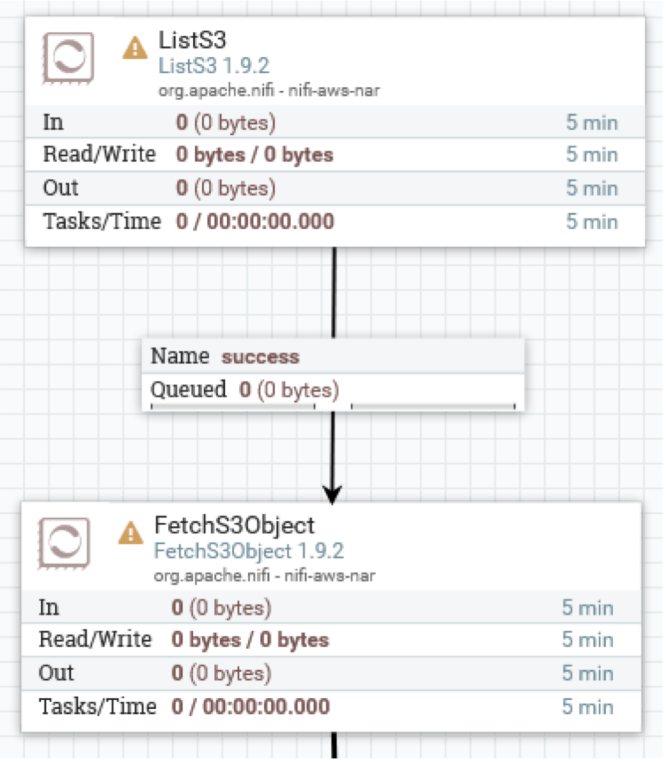
Different categories of processors

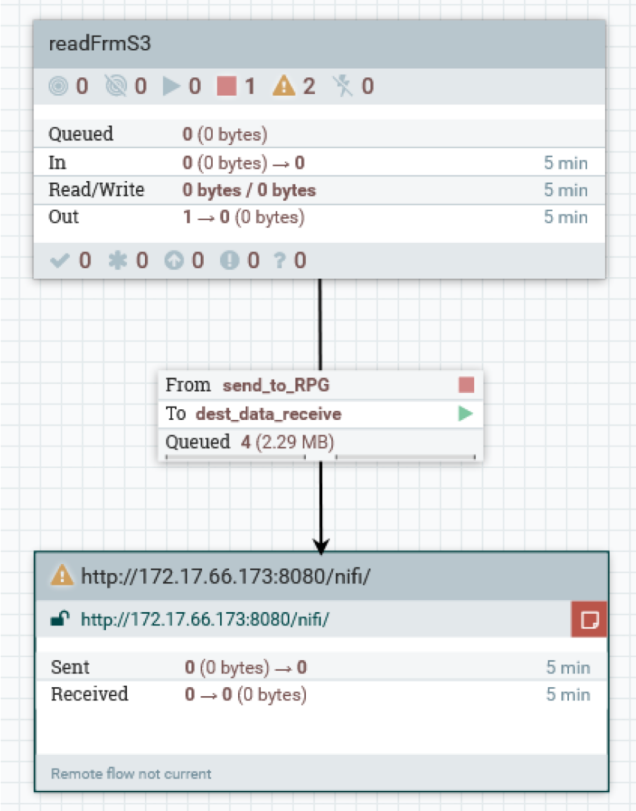
- Data Transformation Processors: ReplaceText, JoltTransformJSON, etc
- Sending Data Processors: PutEmail, PutSFTP, PutFile, PutFTP, etc.
- Splitting and Aggregation Processors: SplitText, SplitJson, SplitXml, MergeContent, SplitContent, etc.
- HTTP Processors: InvokeHTTP , ListenHTTP, etc
- AWS Processors: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

Apache Nifi Data Pipeline

1. Major components

   II. Queue





Apache Nifi Data Pipeline

1. Major components

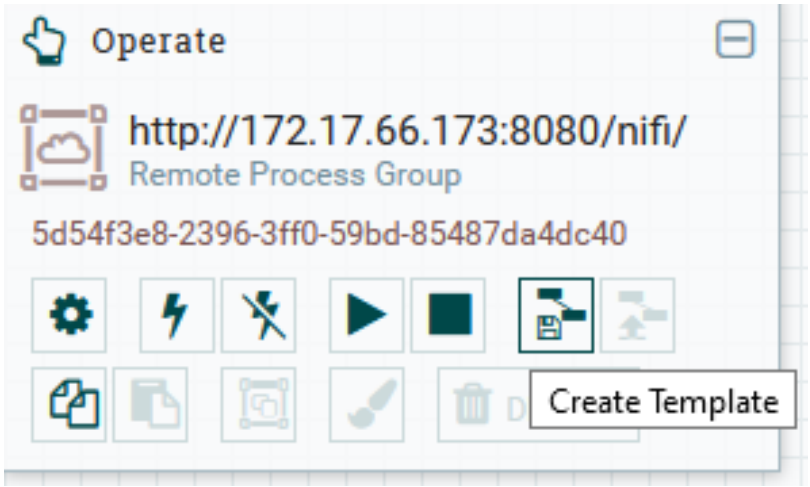
   II. Queue

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

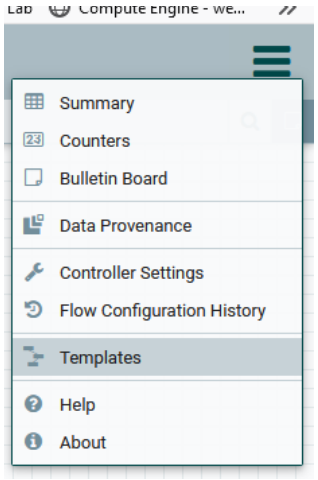
Apache Nifi Data Pipeline

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.



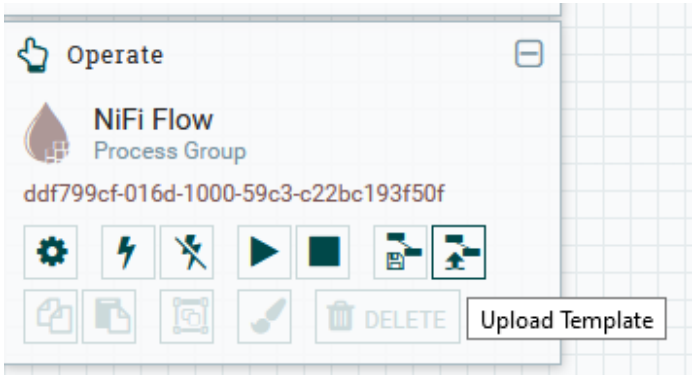
Create Template



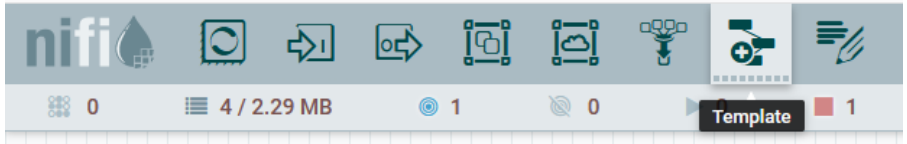
Download Template

Apache Nifi Data Pipeline

Templates:



Upload Template



Add Template

Research on 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

1. https://wiki.oasis-open.org/tosca/TOSCA-implementations
2. Casale, G., Artač, M., van den Heuvel, W. *et al.* RADON: rational decomposition and orchestration for serverless computing. *SICS Softw.-Inensiv. Cyber-Phys. Syst.* (2019). https://doi.org/10.1007/s00450-019-00413-w
3. http://radon-h2020.eu/
4. https://github.com/radon-h2020/radon-particles
5. https://nifi.apache.org/docs/nifi-docs/html/overview.html
6. https://nifi.apache.org/docs.html
7. https://nifi.apache.org/powered-by-nifi.html
8. https://docs.aws.amazon.com/datapipeline/latest/DeveloperGuide/datapipeline-dg.pdf#dp-importexport-ddb-console-start
9. https://www.javatpoint.com/aws-data-pipeline
10. https://aws.amazon.com/datapipeline/
11. https://aws.amazon.com/streaming-data/

Thank you