FIAP

MBA+

MBA EM BIG DATA (DATA SCIENCE)

FIAP

MBA+

INGESTÃO DE DADOS
NiFi

Prof. Thiago Nascimento Nogueira
tnnogueira@gmail.com
Linked-in: https://www.linkedin.com/in/thnogueira

NiFi

Características

FIAP

- “Canivete suiço” para ingestão de dados
- Utilizado em projetos qua não envolvem necessariamente hadoop
- Interface Web
  - Experiência integrada entre desenho, controle, feedback e monitoria
- Altamente configurável
  - Tolerante a falhas com garantia de entrega
  - Baixa latência e alto throughput
  - Priorização dinâmica
  - Fluxo pode ser alterado em tempo de execução
  - “Back pressure”

Características

- Data Provenance (Procedência de dados)
  - Rastreamento do dataflow, do início ao fim
- Desenhado para extensibilidade
  - Permite desenvolvimento de processors customizados
  - Rápido desenvolvimento e testes efetivos
- Segurança
  - SSL, SSH, HTTPS, criptografia de conteúdo, etc..
  - Autorização Multi-tenant e autorização / política de gerenciamento interna

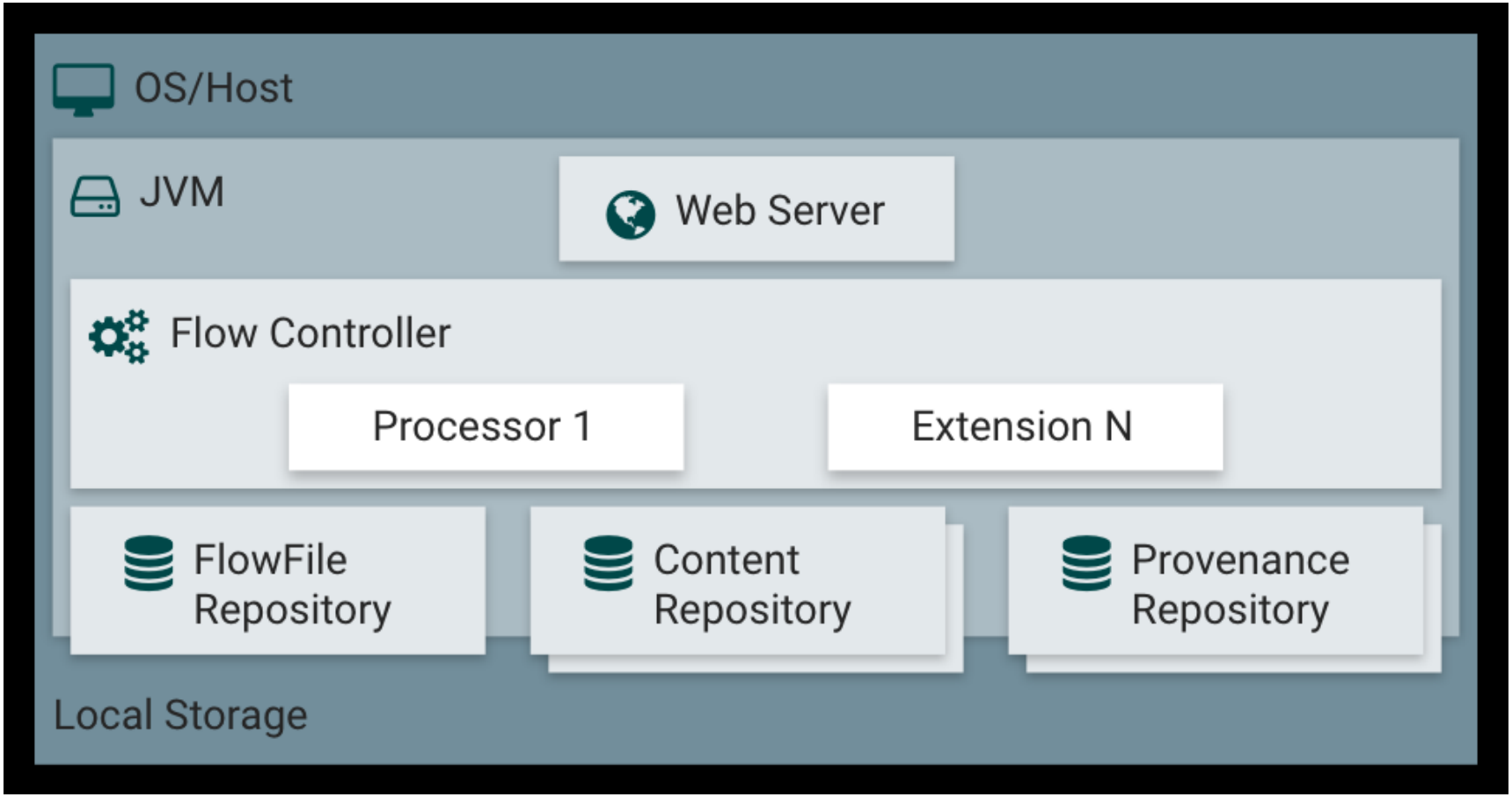
Conceitos Centrais

- **FlowFile** – Objeto que se move através do fluxo onde, para cada FlowFile, o NiFI mantém um mapeamento key/value com atributos e seu conteúdo associado, com zero ou mais bytes
- **Processadores de Fluxo** – Realizam o trabalho de fato. Realizam alguma combinação entre roteamento de dados, transformação, ou mediação entre sistemas. Possuem acesso aos atributos e conteúdos do FlowFile
- **Conexão** – Promovem as ligações entre os processadores. Funcionam como filas e permitem que vários processadores interajam de diferentes maneiras
  - Podem ser priorizadas dinamicamente
  - Possuem limites máximos de carga, permitindo controle tipo “back pressure”
- **Controlador de Fluxo** – Mantém o controle de como os processos se conectam e gerencia os recursos de cada processo
- **Grupo de Processos** – Agrupamento lógico de conjunto de processos.

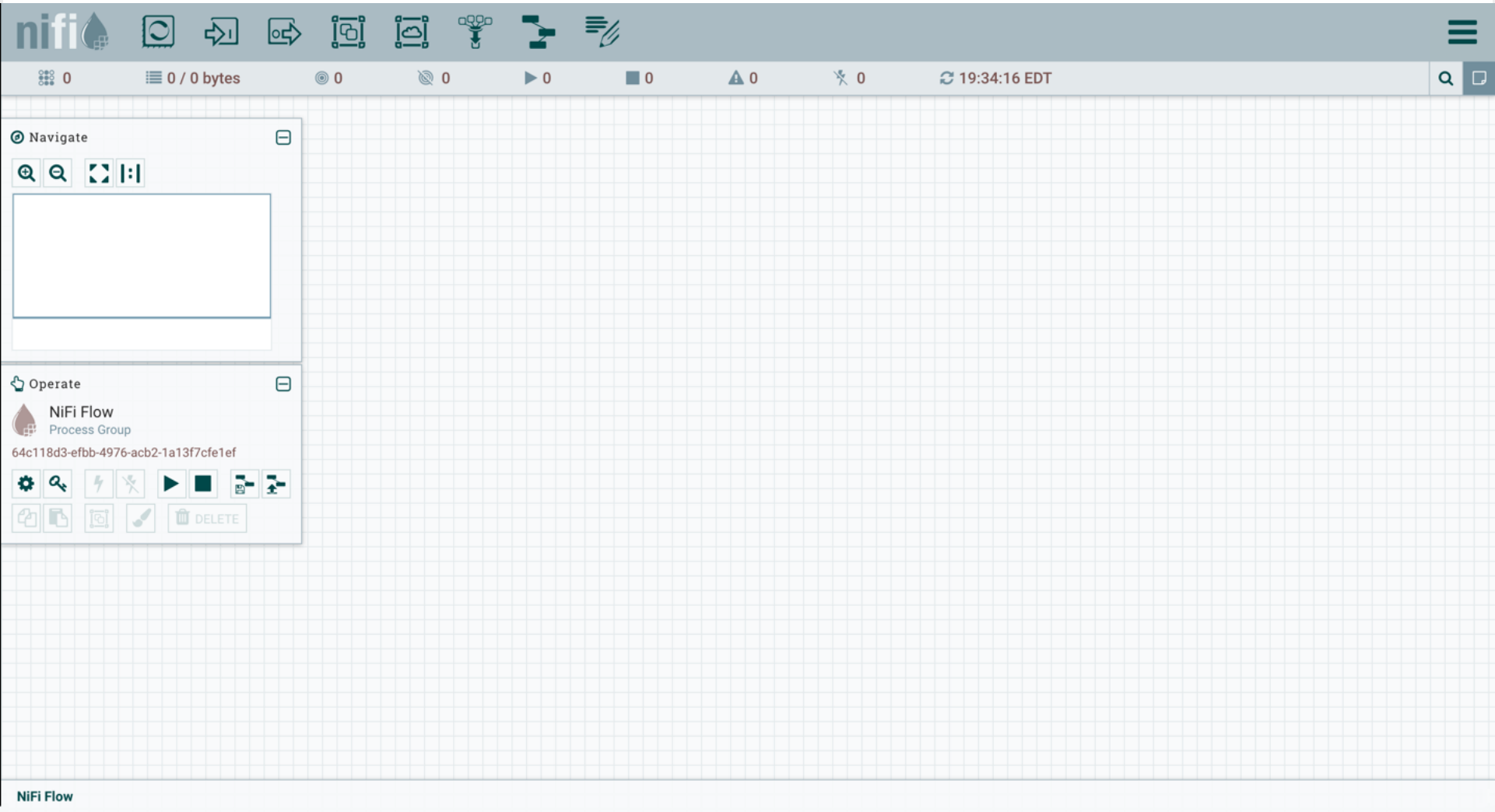
Arquitetura

- **WebServer** – Hospeda os comando NiFi baseados em http e API de controle
- **Flow Controller** – Cérebro da operação. Provê as threads para que os processos rodem e gerencia o agendamento de quando o processo recebe os recursos para executar
- **FlowFile Repository** – Lugar onde o NiFI mantém as informações sobre os estados dos FlowFiles
- Content Repository – Lugar o são armazenados os conteúdos do FlowFile em execução
- **Provenance Repository** – Local onde os dados de eventos de data provenance são armazenados

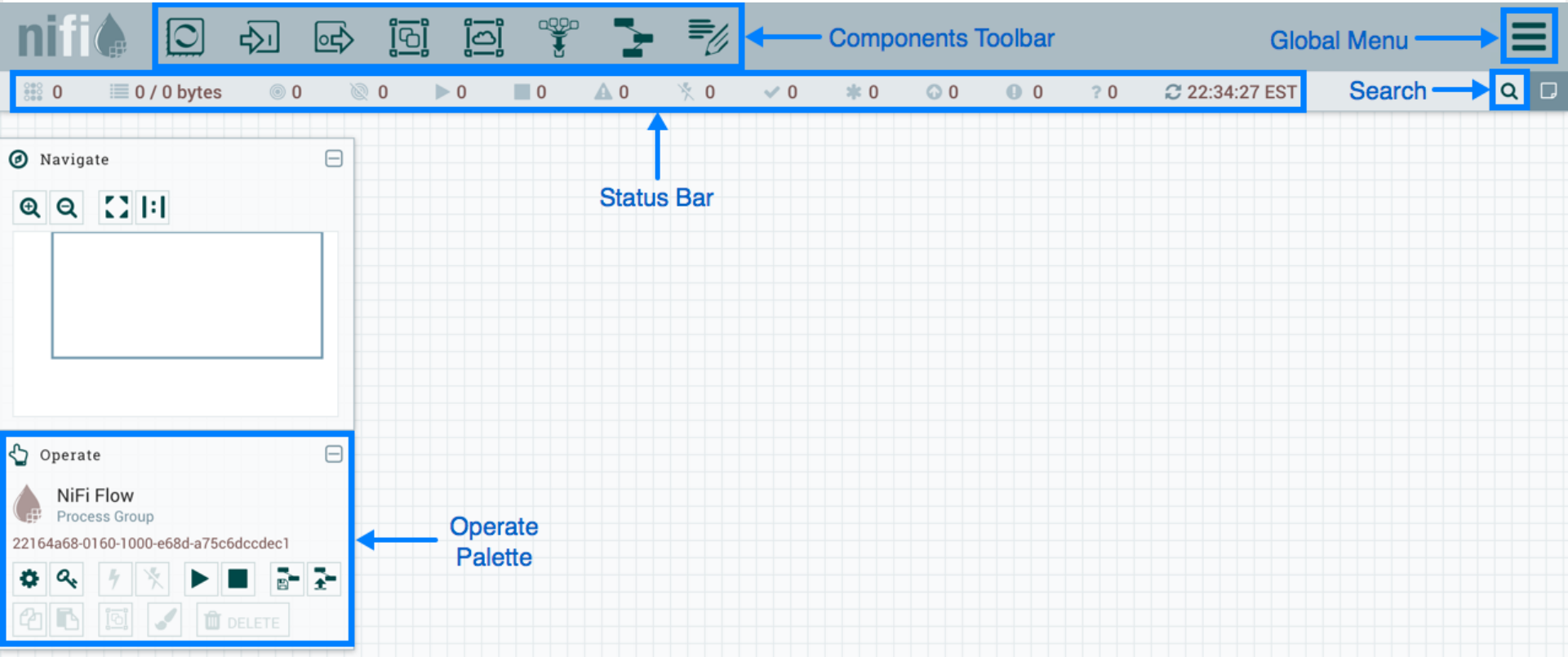
Arquitetura



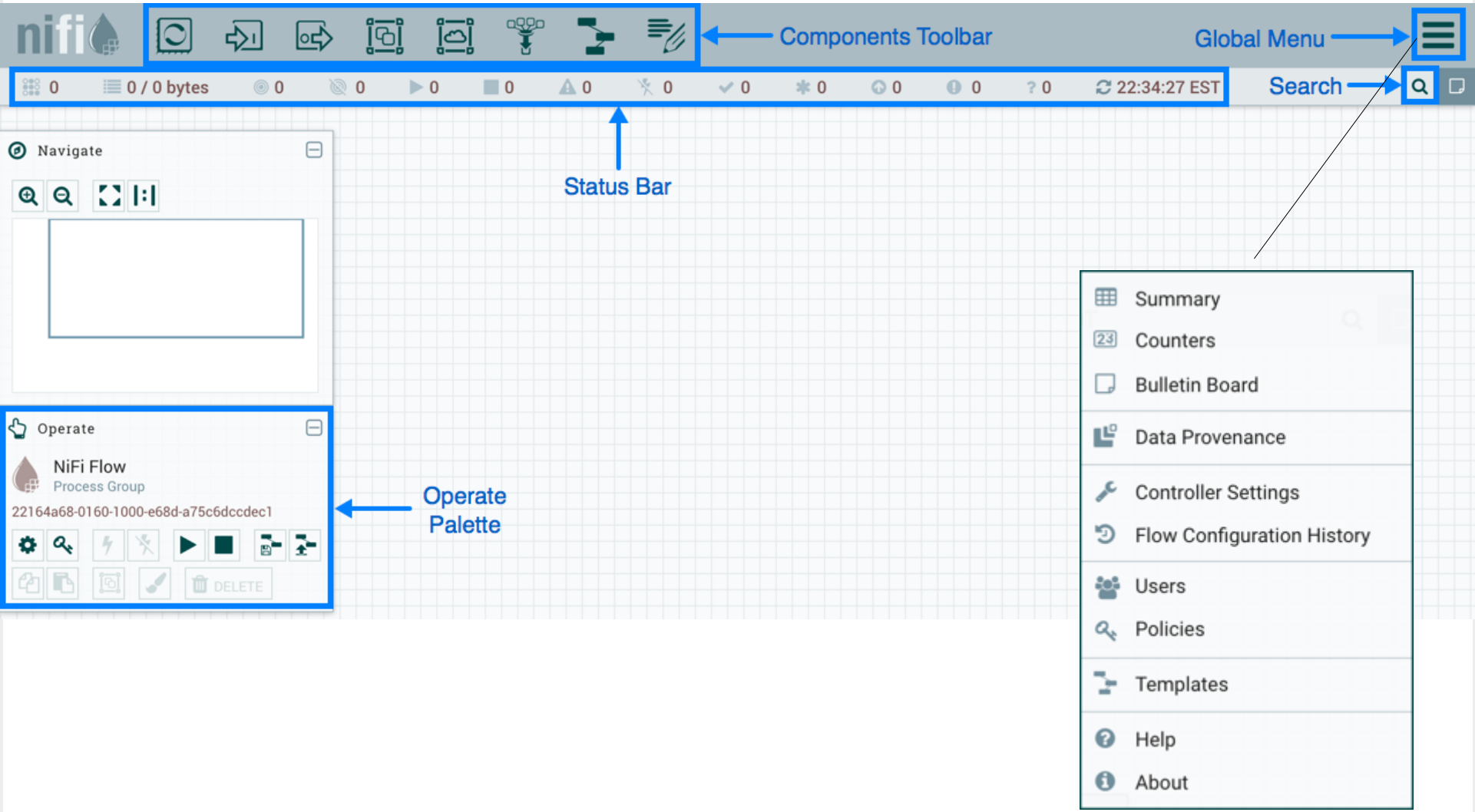
Tela Inicial



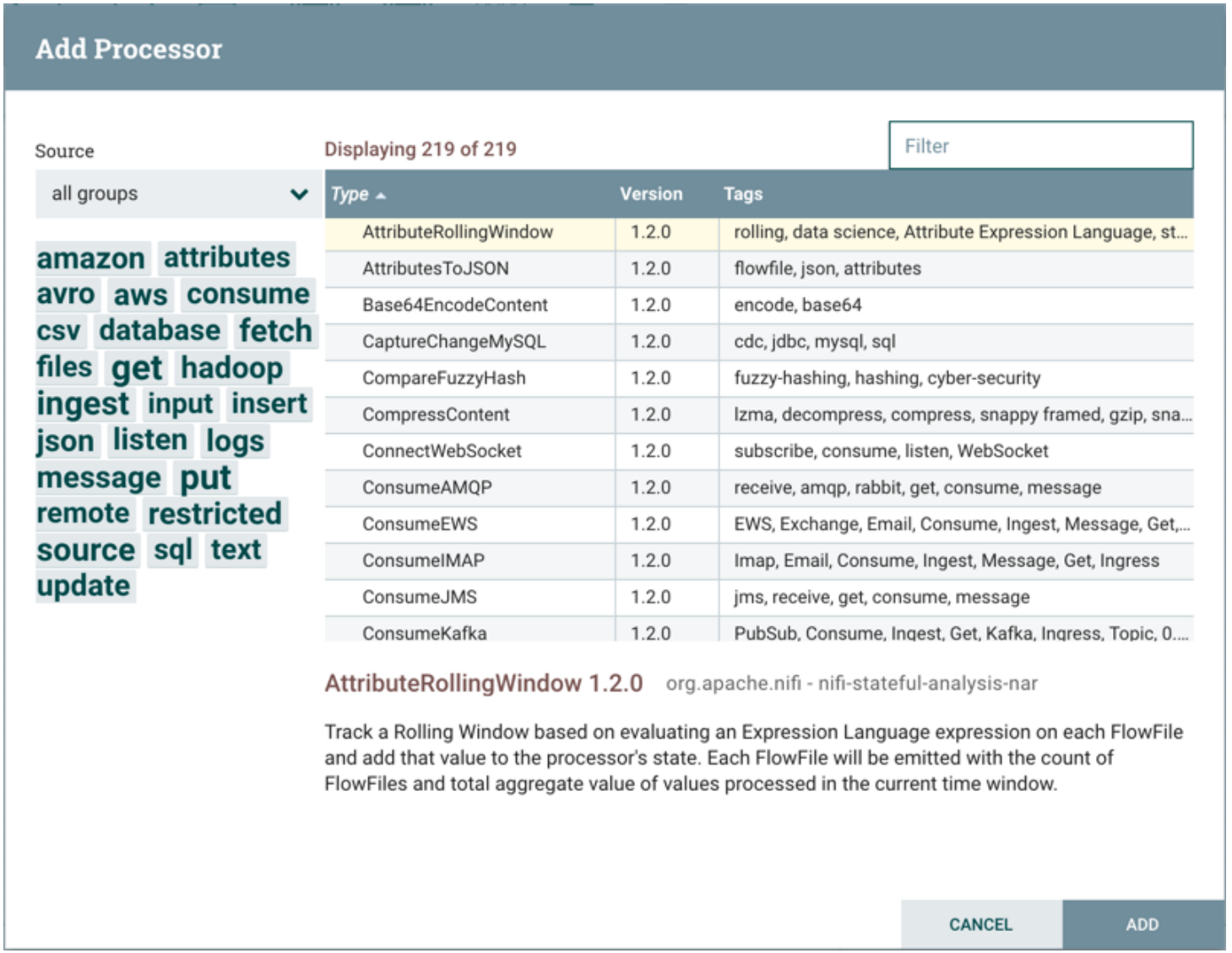
Componentes



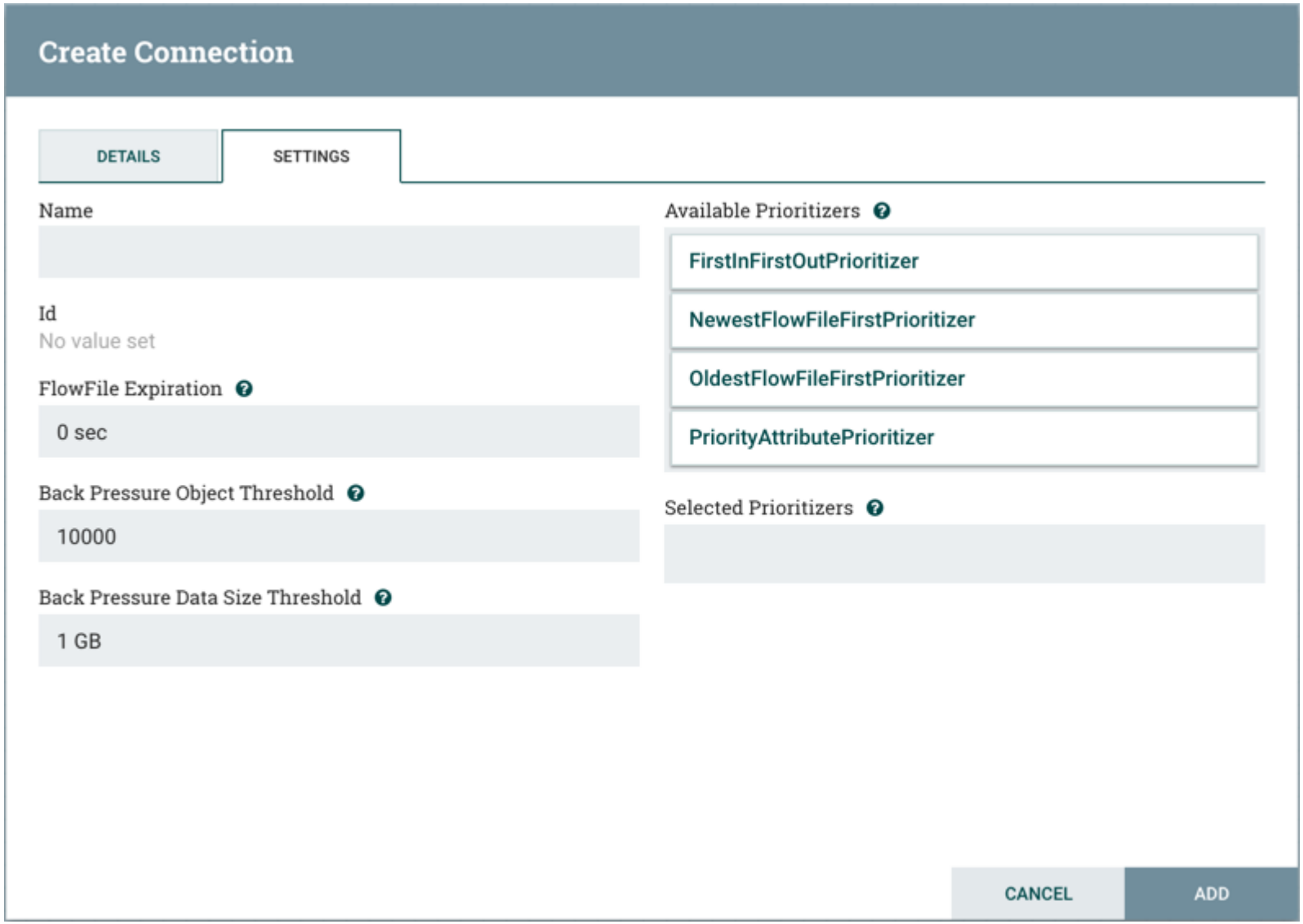
Componentes



Adicionando um Processor

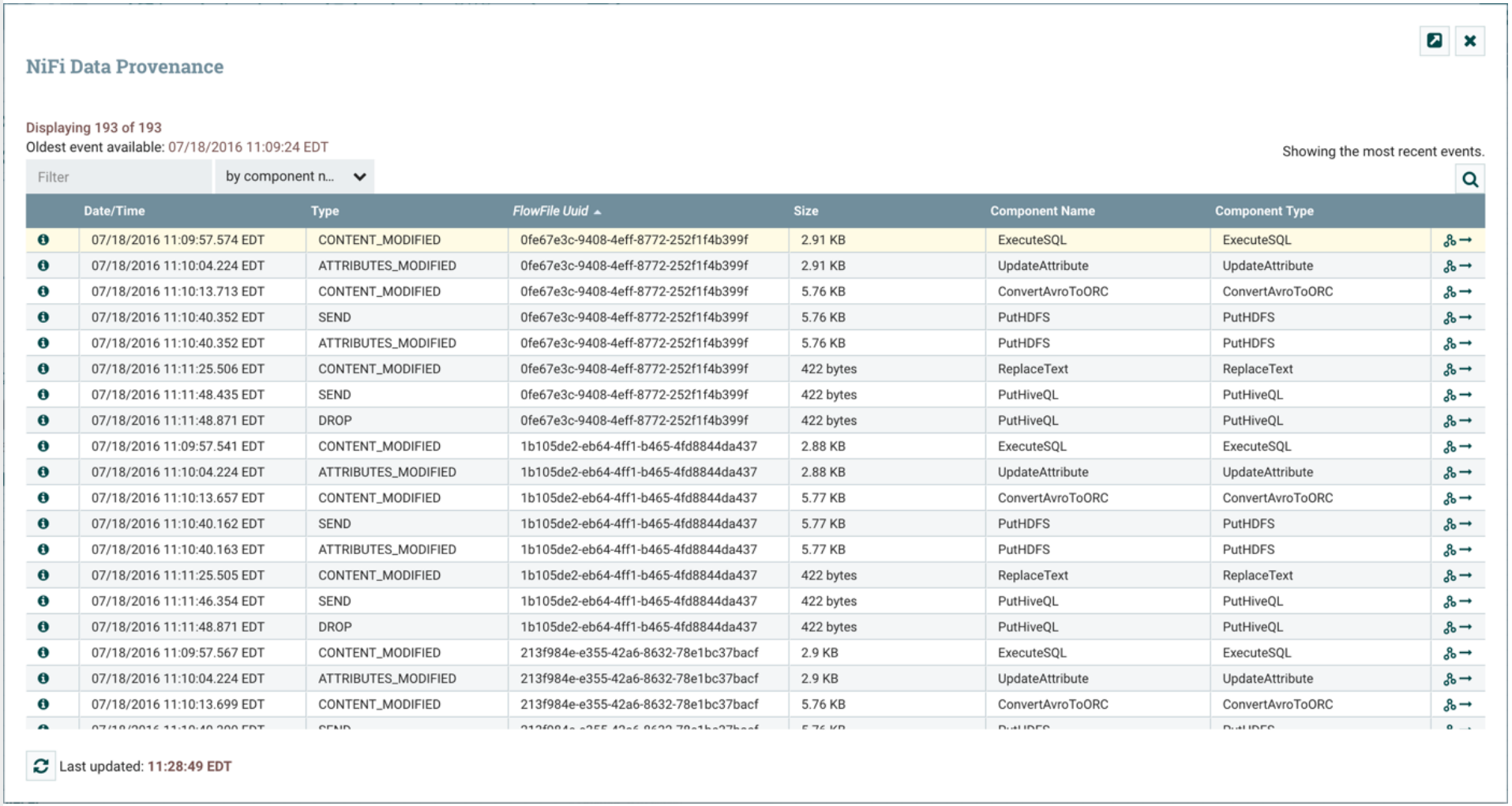


Criando uma conexão



Data Provenance

FIAP



NiFi Data Provenance

Displaying 193 of 193
Oldest event available: 07/18/2016 11:09:24 EDT

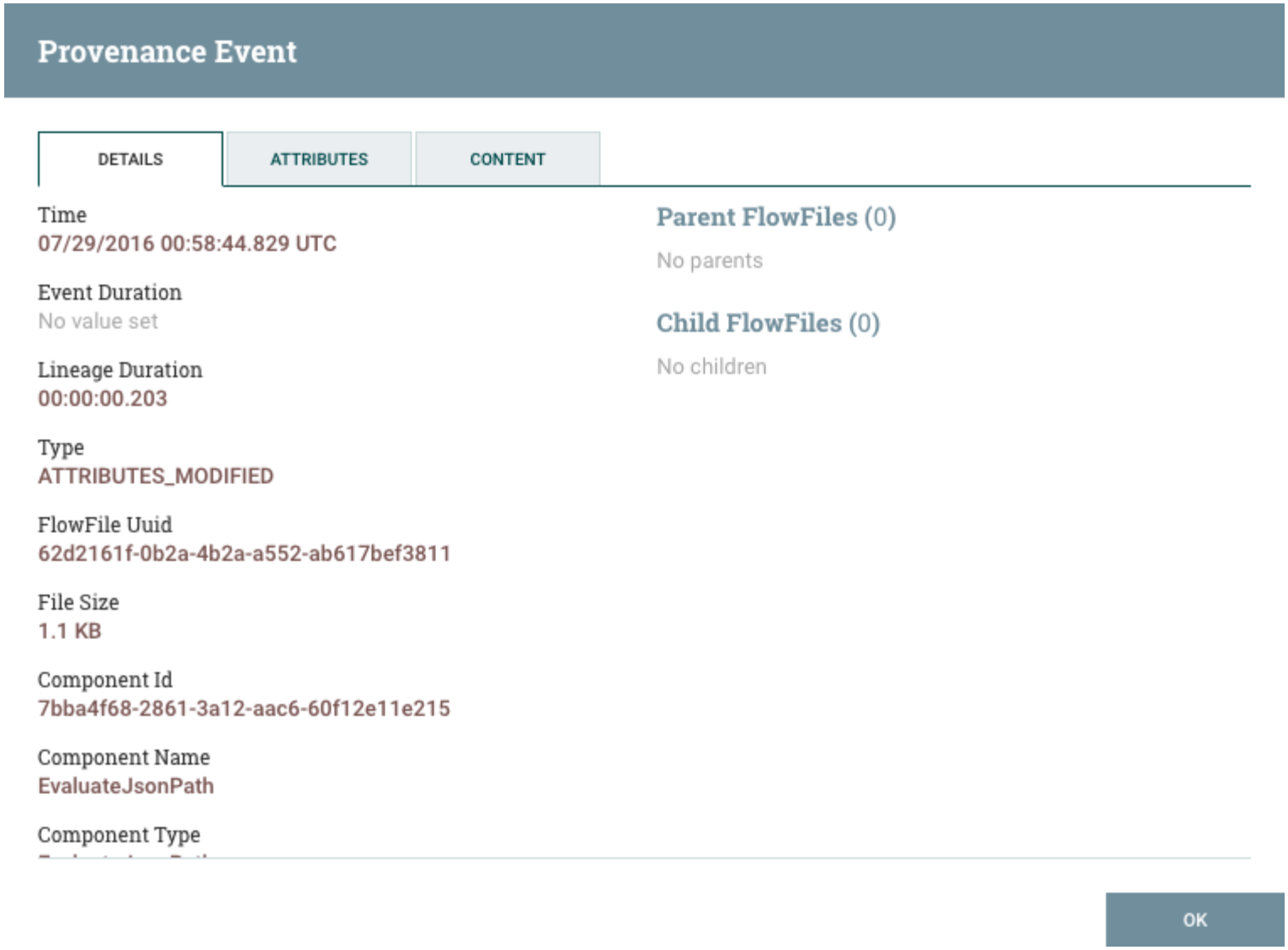
Showing the most recent events.

Filter | by component n...

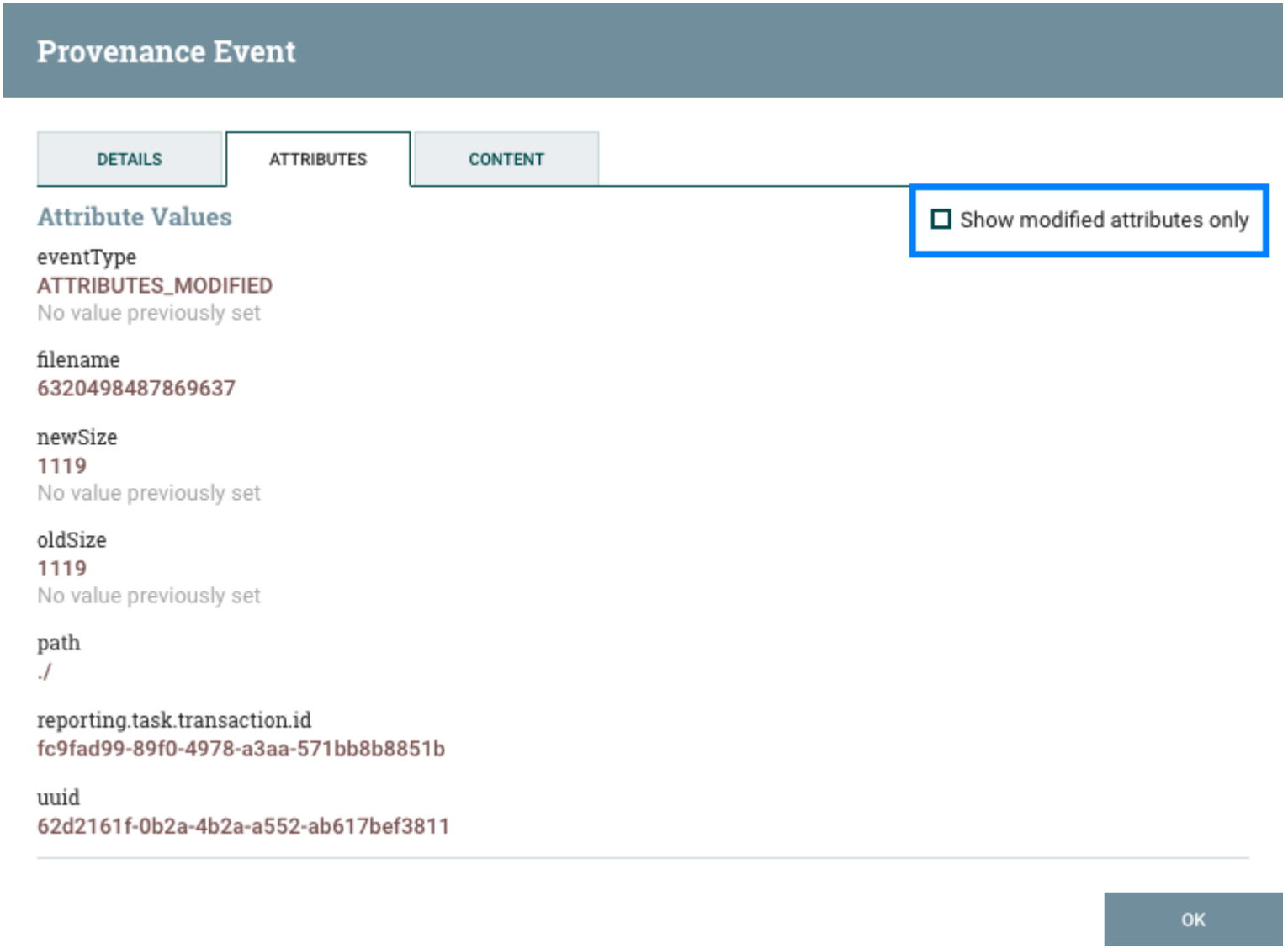
Date/Time	Type	FlowFile Uuid	Size	Component Name	Component Type
07/18/2016 11:09:57.574 EDT	CONTENT_MODIFIED	0fe67e3c-9408-4eff-8772-252f1f4b399f	2.91 KB	ExecuteSQL	ExecuteSQL
07/18/2016 11:10:04.224 EDT	ATTRIBUTES_MODIFIED	0fe67e3c-9408-4eff-8772-252f1f4b399f	2.91 KB	UpdateAttribute	UpdateAttribute
07/18/2016 11:10:13.713 EDT	CONTENT_MODIFIED	0fe67e3c-9408-4eff-8772-252f1f4b399f	5.76 KB	ConvertAvroToORC	ConvertAvroToORC
07/18/2016 11:10:40.352 EDT	SEND	0fe67e3c-9408-4eff-8772-252f1f4b399f	5.76 KB	PutHDFS	PutHDFS
07/18/2016 11:10:40.352 EDT	ATTRIBUTES_MODIFIED	0fe67e3c-9408-4eff-8772-252f1f4b399f	5.76 KB	PutHDFS	PutHDFS
07/18/2016 11:11:25.506 EDT	CONTENT_MODIFIED	0fe67e3c-9408-4eff-8772-252f1f4b399f	422 bytes	ReplaceText	ReplaceText
07/18/2016 11:11:48.435 EDT	SEND	0fe67e3c-9408-4eff-8772-252f1f4b399f	422 bytes	PutHiveQL	PutHiveQL
07/18/2016 11:11:48.871 EDT	DROP	0fe67e3c-9408-4eff-8772-252f1f4b399f	422 bytes	PutHiveQL	PutHiveQL
07/18/2016 11:09:57.541 EDT	CONTENT_MODIFIED	1b105de2-eb64-4ff1-b465-4fd8844da437	2.88 KB	ExecuteSQL	ExecuteSQL
07/18/2016 11:10:04.224 EDT	ATTRIBUTES_MODIFIED	1b105de2-eb64-4ff1-b465-4fd8844da437	2.88 KB	UpdateAttribute	UpdateAttribute
07/18/2016 11:10:13.657 EDT	CONTENT_MODIFIED	1b105de2-eb64-4ff1-b465-4fd8844da437	5.77 KB	ConvertAvroToORC	ConvertAvroToORC
07/18/2016 11:10:40.162 EDT	SEND	1b105de2-eb64-4ff1-b465-4fd8844da437	5.77 KB	PutHDFS	PutHDFS
07/18/2016 11:10:40.163 EDT	ATTRIBUTES_MODIFIED	1b105de2-eb64-4ff1-b465-4fd8844da437	5.77 KB	PutHDFS	PutHDFS
07/18/2016 11:11:25.505 EDT	CONTENT_MODIFIED	1b105de2-eb64-4ff1-b465-4fd8844da437	422 bytes	ReplaceText	ReplaceText
07/18/2016 11:11:46.354 EDT	SEND	1b105de2-eb64-4ff1-b465-4fd8844da437	422 bytes	PutHiveQL	PutHiveQL
07/18/2016 11:11:48.871 EDT	DROP	1b105de2-eb64-4ff1-b465-4fd8844da437	422 bytes	PutHiveQL	PutHiveQL
07/18/2016 11:09:57.567 EDT	CONTENT_MODIFIED	213f984e-e355-42a6-8632-78e1bc37bacf	2.9 KB	ExecuteSQL	ExecuteSQL
07/18/2016 11:10:04.224 EDT	ATTRIBUTES_MODIFIED	213f984e-e355-42a6-8632-78e1bc37bacf	2.9 KB	UpdateAttribute	UpdateAttribute
07/18/2016 11:10:13.699 EDT	CONTENT_MODIFIED	213f984e-e355-42a6-8632-78e1bc37bacf	5.76 KB	ConvertAvroToORC	ConvertAvroToORC

Last updated: 11:28:49 EDT

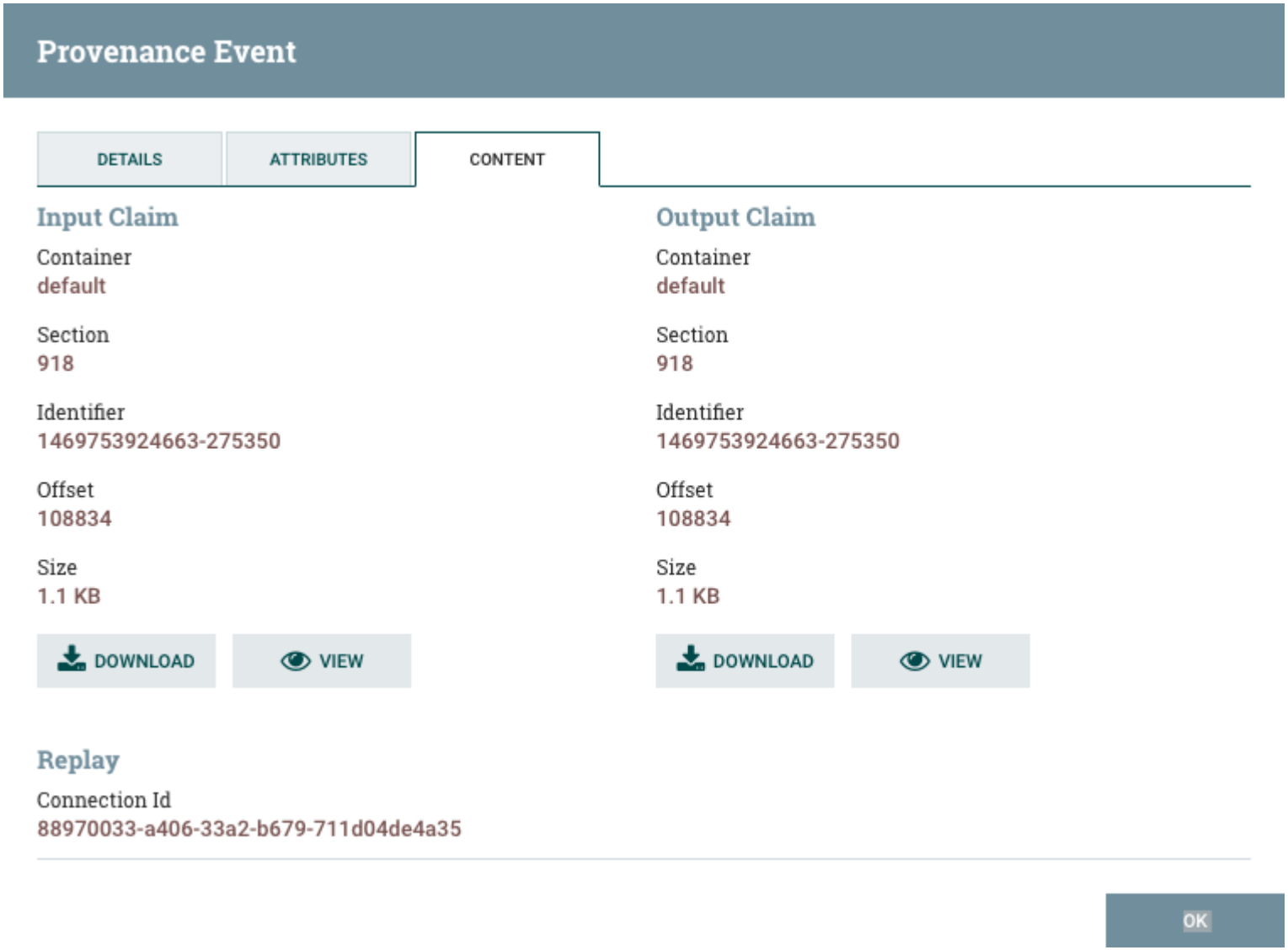
Data Provenance



Data Provenance



Data Provenance



Referências

Gregor Hohpe. Enterprise Integration Patterns [online]. Retrieved: 27 Dec 2014, from: http://www.enterpriseintegrationpatterns.com

Wikipedia. Service Oriented Architecture [online]. Retrieved: 27 Dec 2014, from: http://en.wikipedia.org/wiki/Service-oriented_architecture

Eric Savitz. Welcome to the API Economy [online]. Forbes.com. Retrieved: 27 Dec 2014, from: http://www.forbes.com/sites/ciocentral/2012/08/29/welcome-to-the-api-economy

Adam Duvander. The rise of the API economy and consumer-led ecosystems [online]. thenextweb.com. Retrieved: 27 Dec 2014, from: http://thenextweb.com/dd/2014/03/28/api-economy

Wikipedia. Internet of Things [online]. Retrieved: 27 Dec 2014, from: http://en.wikipedia.org/wiki/Internet_of_Things

Wikipedia. Big Data [online]. Retrieved: 27 Dec 2014, from: http://en.wikipedia.org/wiki/Big_data

Wikipedia. Flow Based Programming [online]. Retrieved: 28 Dec 2014, from: http://en.wikipedia.org/wiki/Flow-based_programming#Concepts

Matt Welsh. Berkeley. SEDA: An Architecture for Well-Conditioned, Scalable Internet Services [online]. Retrieved: 18 Jan 2018, from: http://www.mdw.la/papers/seda-sosp01.pdf