Ambiente Livre
Tecnologia

Hadoop

Treinamento Hadoop – Big Data Open Source - Fundamental.

Instrutor: Marcio Junior Vieira.
marcio@ambientelivre.com.br

O que é PIG

- Apache Pig é uma linguagem de procedimentos de alto nível para consultar grandes conjuntos de dados semiestruturad usando Hadoop e a Plataforma MapReduce

- Pig simplifica o uso de Hadoop permitindo consultas parecidas com SQL em um conjunto de dados distribuídos.
- Pig é uma abstração sobre Hadoop
- Criado pelo Yahoo e hoje é um projeto Apache

PIG Oferece

- Processamento convertido em MapReduce e executado em Clusters Hadoop
-
-

Pig e MapReduce

- **MapReduce**

  - Programadores que raciocinem em funções com o formato mapreduce;

  - Normalmente precisa de programadores java;

- **PIG**

  - Analistas de dados (informatas)

  - Estatísticos

  - Bioinformatas

- Join
- Sort
- Filter
- Data Types
- Group By
- Foreach
- Load

- Order
- Split
- Split
- Store
- Funções definidas pelo Usuário
- Todas Manipulações formais conhecidas do SQL.

Casos comuns

- **ETL**
  - Processar um log de dados
  - Filtrar informações específicas
  - juntar(join) com outros blocos de dados (datasets)
- **Pesquisa em arquivos “RAW”**
  - Auditoria
  - schemas

Usando PIG

- Yahoo (de 40% a 60% das cargas de trabalho do Hadoop são geradas de scripts do Pig Latin)
- Twitter (processando logs, minerando dados de tweet)
- Netfix
- Linkedin (usado para descobrir pessoas que possa conhecer)
- Ebay (usando o Pig para otimização de procura)
- MapQuest (análises e processamento de dados em lote)
- AOL

Componentes

- **PIG Latin**
  - Linguagem baseada em comandos
  - Desenhada específicamente para controle de fluxo e transformação (etl)
- **Ambiente de Execução**
  - Existem dois modos de execução (Local e Hadoop)
- **Compilador converte o PIG em MapReduce**
  - Assim o compilador otimiza as execuções
  - Atualizações do PIG podem conter melhorias

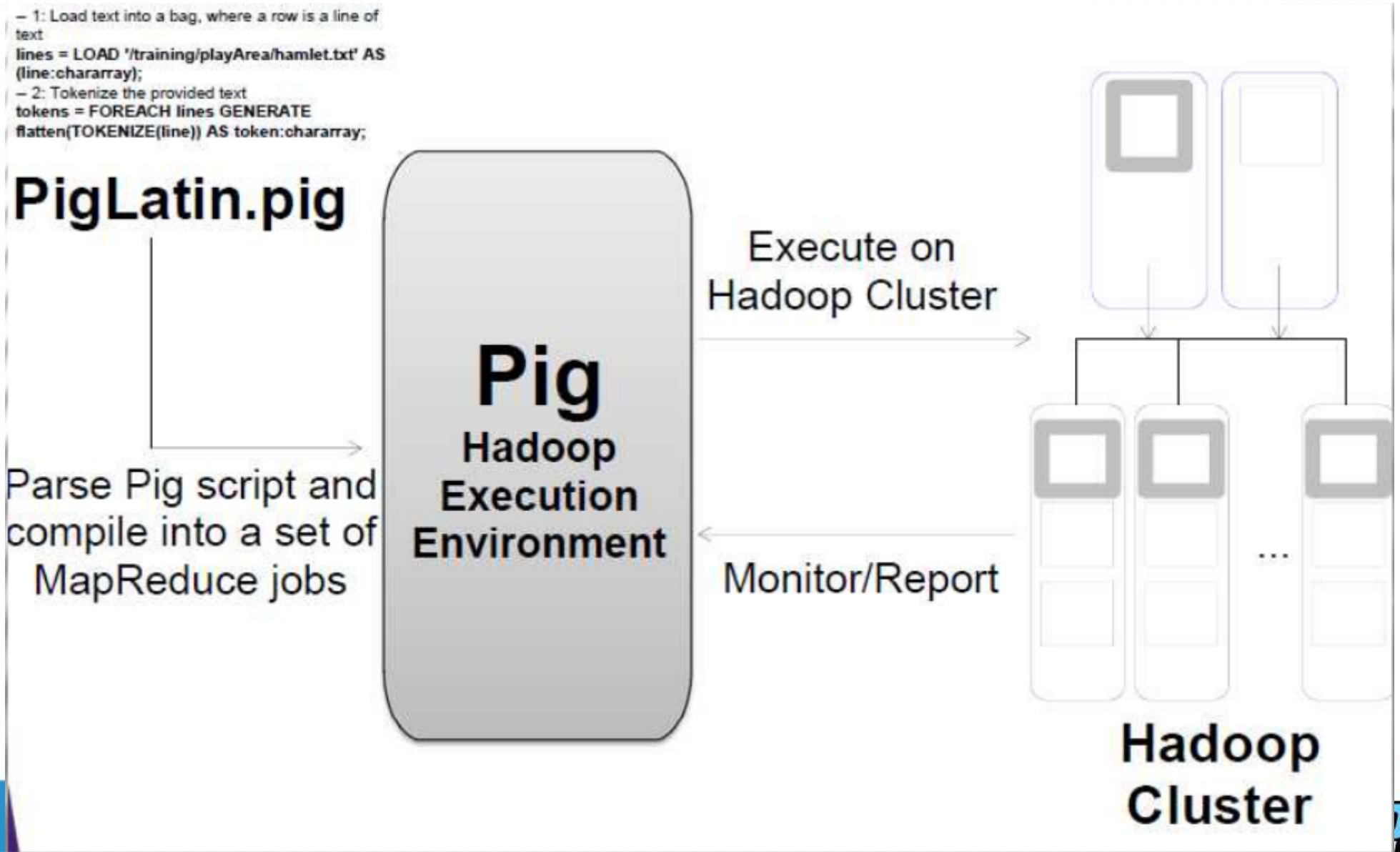
Modos de Execução

- **Local**
  - Executa em um única JVM
  - trabalha exclusivamente no sistema de arquivos local.
  - excelente para desenvolvimento, experimentação e protótipos
- **Hadoop**
  - Também conhecido como MapReduce
  - o PIG adapta o Latin dentro das tarefa do MapReduce e, assim são executadas nos clusters

  - Pode ser usado nas instalações do Hadoop semi-distribuído ou totalmente distribuído

Ambiente Livre
Tecnologia

Modo Hadoop



Instalação

- Java 1.6 ou Superior
- **Configurar variável**
  $JAVA_HOME
  PIG_HOME=$CDH_HOME/pig-0.9.2-cdh4.0.0
  PATH=$PATH:$PIG_HOME/bin
  $HADOOP_HOME
  $HADOOP_CONF_DIR
- Em Linux ou Cygwin on Windows
- **Help;**

  - pig -help
-
-

Opções -x

- $pig -x local
- $pig -x mapreduce
-

Modo de Execução

- **Script**
  $ pig scriptFile.pig
- **Grunt**

  - Shell interativo
- **Embarcado (embedded)**
  - Executa os comandos pig usando o PigServer class (Da mesma forma que o JDBC executa SQL)
  - Isso permita, também acessar via código, no modo Grunt, via PigRunner Class

Conceitos de PigLatin

- **Campo** (field): uma peça de dados
- **Tupla** (tuple): conjunto ordenado de campos, representados entre “(“ e “)”
  Ex: (10.4, 5, word, 4 , field1)
- **Bag** coleção (collection) de tuplas, representadas entre “{“ e “}”
  Ex. {(10.4, 5, word, 4 , field1), (this, 1, hahaha, hello world Z) }
- **Analogia**

  - Bag -> tabela

  - Tupla-> linha
- **Características**
  - As tuplas não precisam ter o mesmo número de campos em um Bag

Pig Latin

```
$ pig
grunt> cat /training/playArea/pig/a.txt
a 1
d 4
c 9
k 6
grunt> records = LOAD '/training/playArea/pig/a.txt' as
(letter:chararray, count:int);
grunt> dump records;
...
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher -
50% complete
2012-07-14 17:36:22,040 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher -
100% complete
...
(a,1)
(d,4)
(c,9)
(k,6)
grunt>
```

Ambiente Livre
Tecnologia

Pig Latin

```
grunt> chars = LOAD '/training/playArea/pig/b.txt' AS (c:chararray);
grunt> describe chars;
chars: {c: chararray}
grunt> dump chars;
(a)
(k)
...
...
(k)
(c)
(k)
grunt> charGroup = GROUP chars by c;
grunt> describe charGroup;
charGroup: {group: chararray,chars: {(c: chararray)}}
grunt> dump charGroup;
(a,{(a),(a),(a)})
(c,{(c),(c)})
(i,{(i),(i),(i)})
(k,{(k),(k),(k),(k)})
(l,{(l),(l)})
```

Grande Volume de Dados

- Normalmente o ambiente Hadoop trabalha com um grande volume de dados, assim não faz sentido imprimir tudo em tela. Assim é comum emitir os resultados com o comando STORE, ao invés de DUMP;
- Para fins de analise e depuração, é interessante limitar a saída para um pequeno subset de dados para a tela;

Tipos de Dados

Type	Description	Example
	Simple	
int	Signed 32-bit integer	10
long	Signed 64-bit integer	10L or 10l
float	32-bit floating point	10.5F or 10.5f
double	64-bit floating point	10.5 or 10.5e2 or 10.5E2
	Arrays	
chararray	Character array (string) in Unicode UTF-8	hello world
bytearray	Byte array (blob)	
	Complex Data Types	
tuple	An ordered set of fields	(19,2)
bag	An collection of tuples	{(19,2), (18,1)}
map	An collection of tuples	[open#apache]

Source: Apache Pig Documentation 0.9.2; "Pig Latin Basics". 2012

Referências

- http://pig.apache.org
-