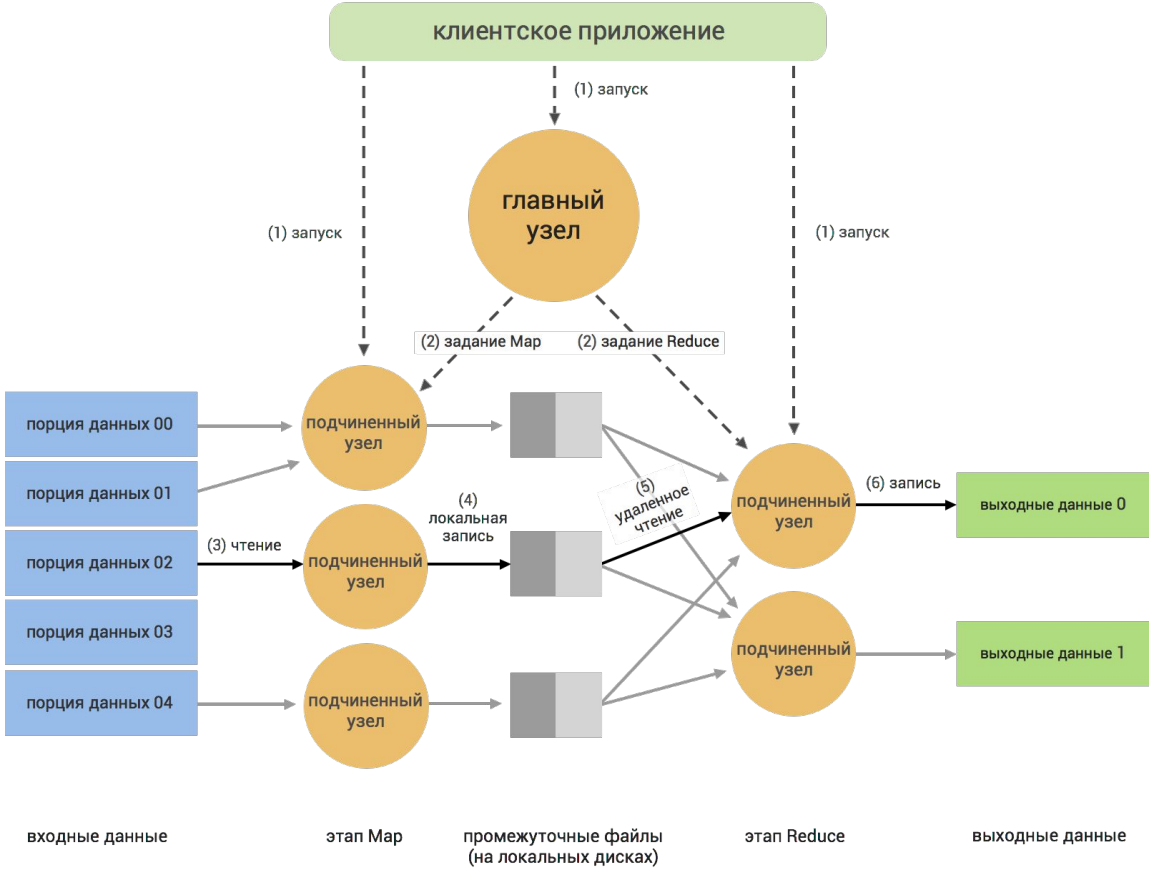
PIG


клиентское приложение
(1) запуск
(1) запуск
(1) запуск
главный узел
(2) задание Map
(2) задание Reduce
порция данных 00
порция данных 01
порция данных 02
порция данных 03
порция данных 04
подчиненный узел
подчиненный узел
подчиненный узел
(3) чтение
(4) локальная запись
(5) удаленное чтение
подчиненный узел
подчиненный узел
(6) запись
выходные данные 0
выходные данные 1
входные данные
этап Map
промежуточные файлы (на локальных дисках)
этап Reduce
выходные данные

Hello!

I am PIG

I am working in Yahoo!

Особенности:

- ✗ язык Pig Latin
- ✗ Интерактивная консоль
- ✗ Встроенные функции агрегации
- ✗ Поддержка пользовательских функций (UDF, User-defined function)
- ✗ Данные — в виде структур (Tuple, Bag)
- ✗ с текстовыми файлами (можно задать разграничительный символ)
- ✗ с сжатыми текстовыми файлами (Gzip, Bzip)
- ✗ имеет огромное количество встроенных функций для работы с: датами, строками, структурами
- ✗ с математическими функциями
- ✗ Если всего перечисленного выше не хватило, то можно использовать кастомные функции (jython, java)

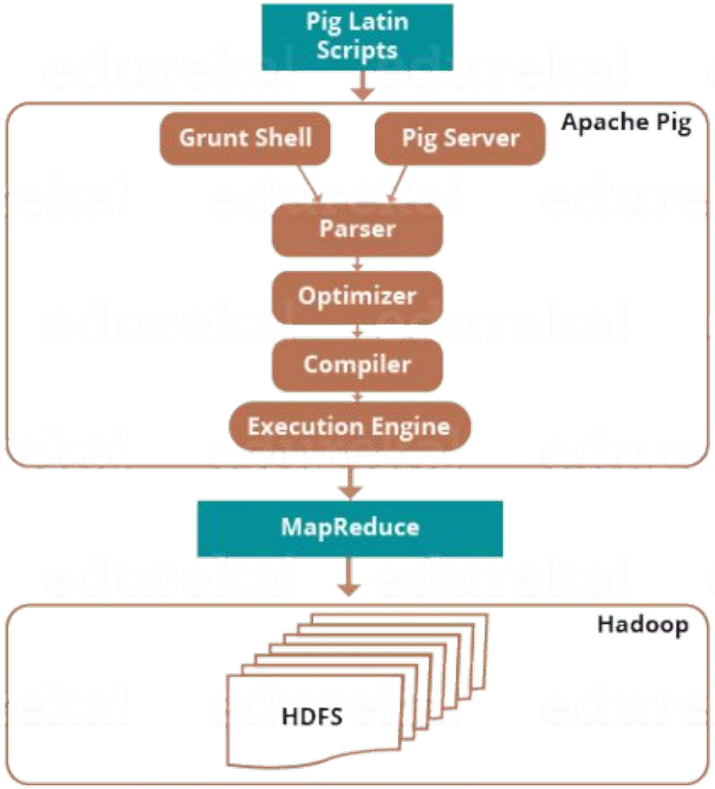
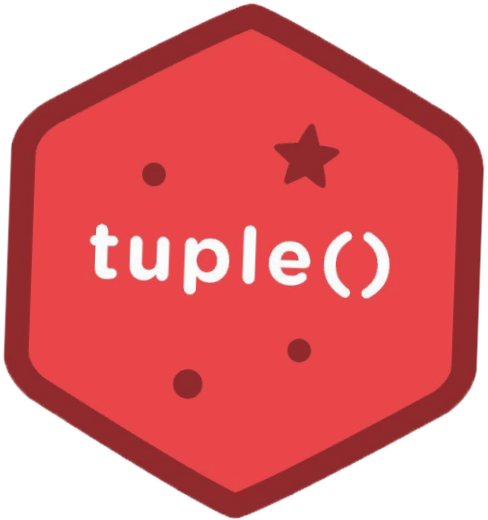


Figure: Apache Pig Architecture

Tuple

- ✗ Упорядоченный набор полей. Структура, к полям которой можно обращаться по индексу и/или имени.



```
time,  bid_id,  user_id,  dsp_id,  bid
------------------------------------------------------------------
----
(2014.02.14 14:08:27.711,  56949,  45234534553459,  DSP-2,  12)
(2014.02.14 14:08:28.712,  61336,  45221696259999,  DSP-1,  56)
(2014.02.14 14:08:29.713,  74685,  45221699381039,  DSP-2,  89)
(2014.02.14 14:08:30.714,  56949,  45221695781716,  DSP-1,  21)
(2014.02.14 14:08:25.715,  27617,  45221682863705,  DSP-3,  22)
```

Bag

- Коллекция (множество) Tuple.

```
time,  bid_id,  user_id,  dsp_id,  bid
-------------------------------------------------------------------
----
{(2014.02.14 14:08:27.711,  56949,  45234534553459,  DSP-2,  12)
 (2014.02.14 14:08:28.712,  61336,  45221696259999,  DSP-1,  56)
 (2014.02.14 14:08:29.713,  74685,  45221699381039,  DSP-2,  89)
 (2014.02.14 14:08:30.714,  56949,  45221695781716,  DSP-1,  21)
 (2014.02.14 14:08:25.715,  27617,  45221682863705,  DSP-3,  22)}
```

Базовые функции:

- LOAD <путь_файла>
  USING PIGSTORAGE ('<знак_разделения>')
  AS (<колонка_1> : <тип>, ...);
- STORE <название_таблицы>
  INTO <путь_файла>;
- FOREACH <таблица_1> GENERATE <условия>;
- JOIN <таблица_1> BY <параметр>,
  <таблица_2> BY <параметр>;
- GROUP <таблица> BY <параметр>;
- FILTER <таблица> BY <параметр>;

Базовые функции:

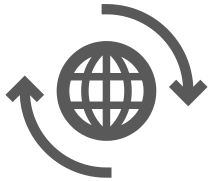
- ✗ UNION <запрос_1>, <запрос_2>;
- ✗ DISTINCT <таблица>;
- ✗ ORDER <таблица> BY <параметр> (ASK / DESC);
- ✗ SPLIT <таблица> INTO
  <нов_таблица_1> IF <условие>,
  <нов_таблица_2> IF <условие>;

Типы данных

Pig Data type	Implementing Class
Bag	org.apache.pig.data.DataBag
Tuple	org.apache.pig.data.Tuple
Map	java.util.Map<Object, Object>
Integer	java.lang.Integer
Long	java.lang.Long
Float	java.lang.Float
Double	java.lang.Double
Chararray	java.lang.String
Bytearray	byte[]

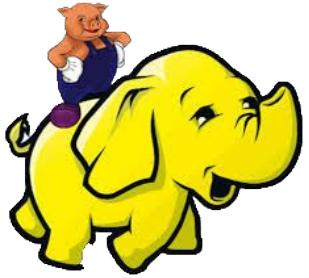
Extract, Transform, Load. (ETL)

```
records = LOAD '/log/flume/events/14-02-20/'
USING PigStorage('\t')
AS (
date:chararray,
clientip:chararray,
clientport:chararray,
proto:chararray,
statuscode:int,
bytes:int,
sq:chararray,
bq:chararray,
request:chararray );
```



Extract

Transform



- count_total = FOREACH (GROUP records ALL) GENERATE COUNT(records);
- count_ip = FOREACH (GROUP records BY clientip) GENERATE group AS ip, COUNT(records) AS cnt;
- top_ip = ORDER count_ip BY cnt DESC;

Load

```
%declare DT `date +%y%m%dT%H%M` STORE
count_total INTO '$DT/count_total'; STORE top_ip
INTO '$DT/top_ip';
STORE top_req INTO '$DT/top_req';
```

Pig Механизмы исполнения

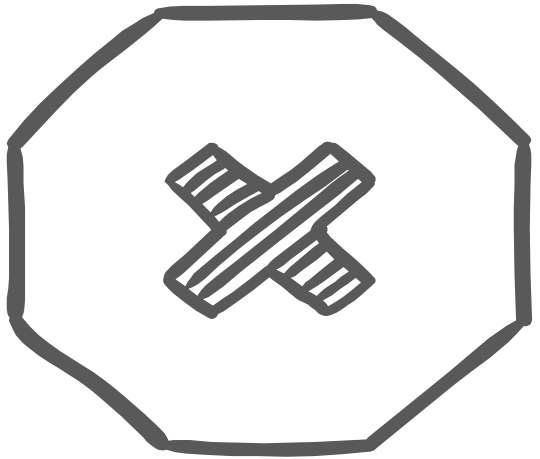
- Интерактивный режим(оболочка Grunt)
- Пакетный режим (скрипт)
- Встроенный режим (UDF)

Операторы диагностики

Оператор	Описание
DESCRIBE	Возвращает схему массива.
DUMP	Выводит содержимое массива на экран.
EXPLAIN	Показывает планы исполнения MapReduce.

- ✗ Процедурный подход.
- ✗ Формирование MapReduce.
- ✗ Интерактивность.
- ✗ Быстрота разработки.

- ✗ Не всё укладывается в Pig
- ✗ Pig Latin более сложен
- ✗ Для UDF используется Java




YAHOO!


mapquest


ADYARD
اديـــارد


LinkedIn


Aol.


twitter


ebay

Thanks!

Any questions?