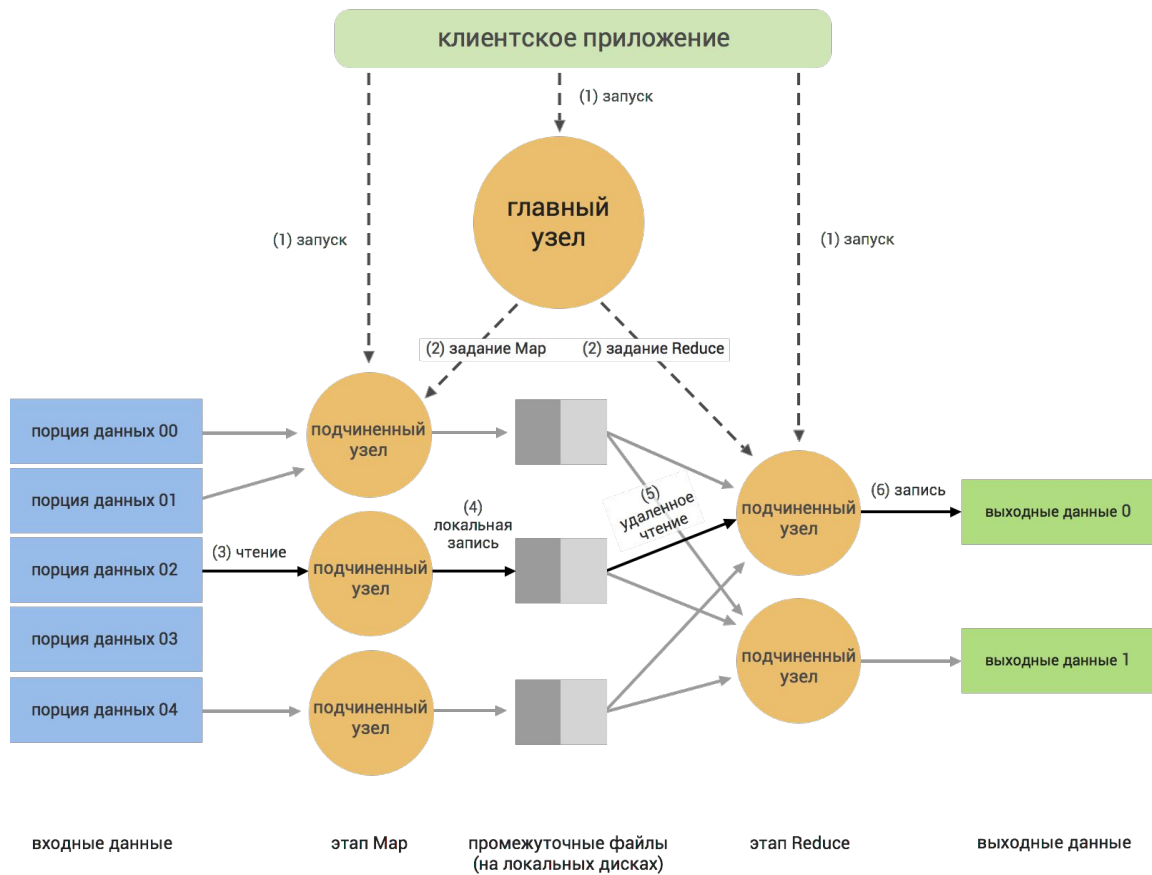




PIG



Hello!

I am PIG

I am working in Yahoo!



Особенности:

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

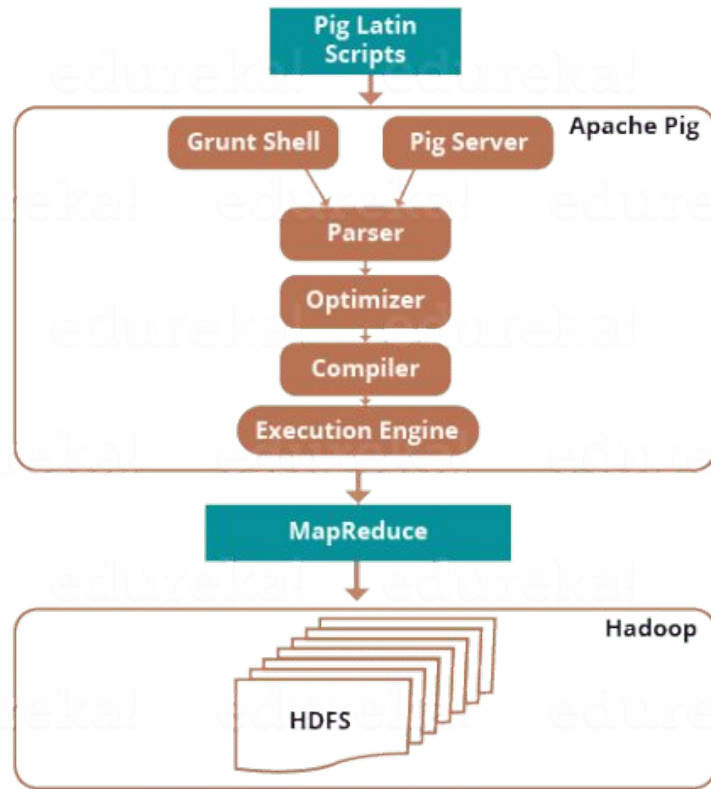
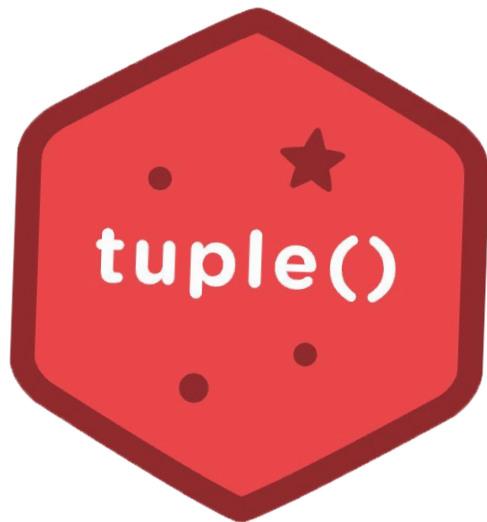


Figure: Apache Pig Architecture

Tuple

- X 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

X 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:

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

$f(x)$

Базовые функции:

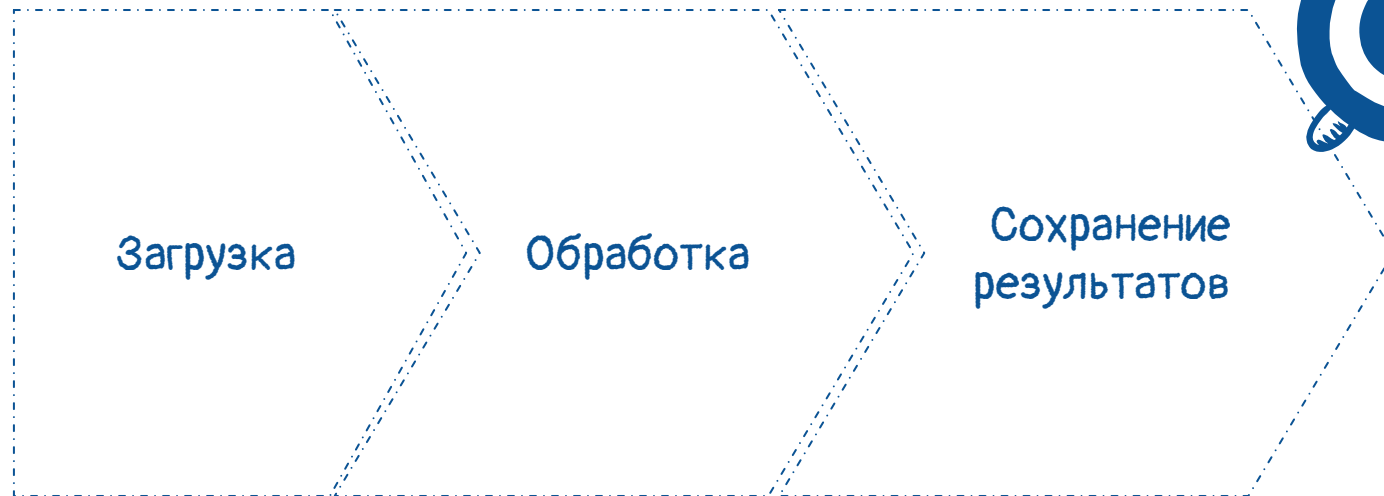
- X UNION <запрос_1>, <запрос_2>;
- X DISTINCT <таблица>;
- X ORDER <таблица> BY <параметр> (ASC / DESC);
- X SPLIT <таблица> INTO
 <нов_таблица_1> IF <условие>,
 <нов_таблица_2> IF <условие>;

$f(x)$

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)



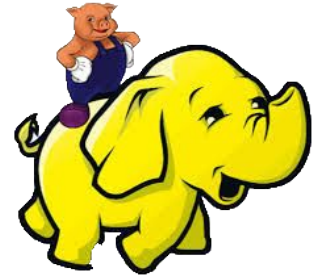
```
records = LOAD '/log/flume/events/14-02-20/'  
USING PigStorage('\t')  
AS (  
date:chararray,  
clientip:chararray,  
clientport:chararray,  
proto:chararray,  
statuscode:int,  
bytes:int,  
sq:chararray,  
bq:chararray,  
request:chararray );
```



Extract



Transform



- X** `count_total = FOREACH (GROUP records ALL) GENERATE COUNT(records);`
- X** `count_ip = FOREACH (GROUP records BY clientip) GENERATE group AS ip, COUNT(records) AS cnt;`
- X** `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';
```



Риг Механизмы исполнения

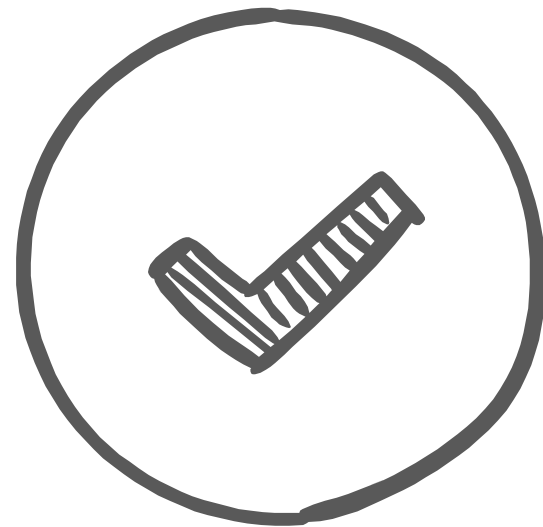
- х Интерактивный режим (оболочка Grunt)
- х Пакетный режим (скрипт)
- х Встроенный режим (UDF)



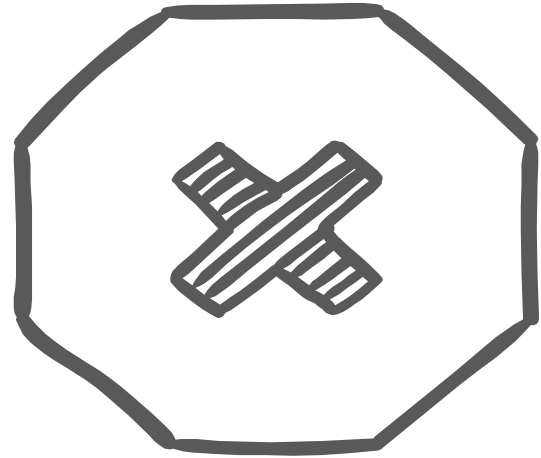
Операторы диагностики

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

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



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



YAHOO!



Aol.



ebay™

Thanks!

Any questions?

