



Резюме модуля

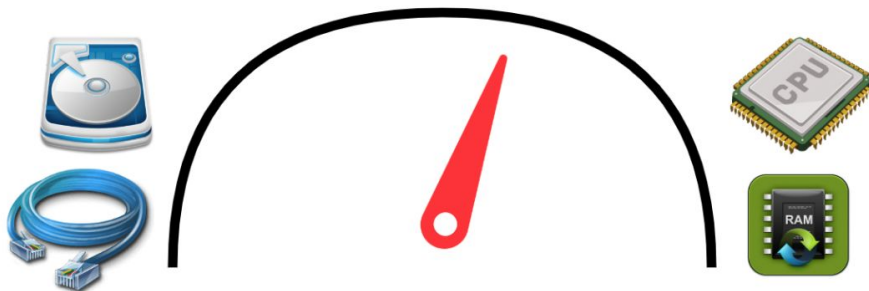
Драль Алексей, study@bigdatateam.org

CEO at BigData Team, <https://bigdatateam.org>

<https://www.facebook.com/bigdatateam>




- ✓ Кодирование vs Сжатие
- ✓ “горячие” vs “холодные” данные





- ✓ Кодирование vs Сжатие
- ✓ “горячие” vs “холодные” данные
- ✓ Hive: File vs Row format

```
CREATE EXTERNAL TABLE tab_dataset (  
    first_column    STRING,  
    second_column   STRING,  
    value           INT  
)  
ROW FORMAT DELIMITED  
    FIELDS TERMINATED BY '\001'  
    COLLECTION ITEMS TERMINATED BY '\002'  
    MAP KEYS TERMINATED BY '\003'  
    LINES TERMINATED BY '\n'  
STORED AS file_format  
LOCATION '/user/<user>/hive_practice_data/';  
  
default: TEXTFILE
```





- ✓ Кодирование vs Сжатие
- ✓ “горячие” vs “холодные” данные
- ✓ Hive: File vs Row format
- ✓ RCFile vs ORC vs Parquet

Логическая структура

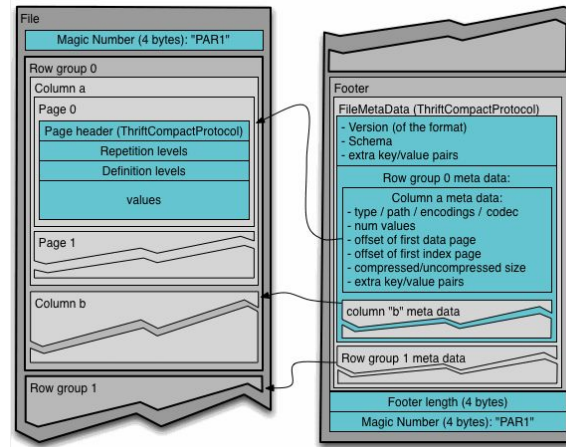
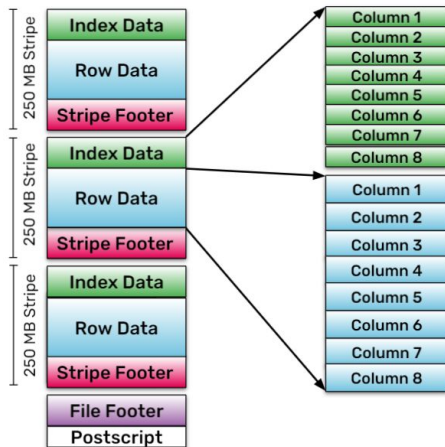
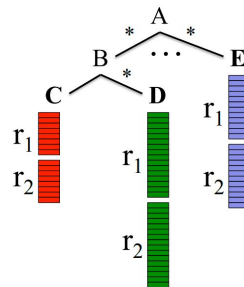
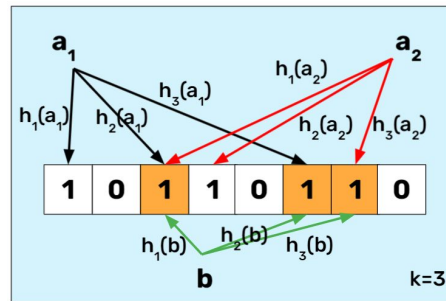
	col1	col2	col3
row1	1	2	3
row2	4	5	6
row3	7	8	9
row4	10	11	12

Row-oriented layout

row1	row2	row3	row4
1 2 3	4 5 6	7 8 9	10 11 12

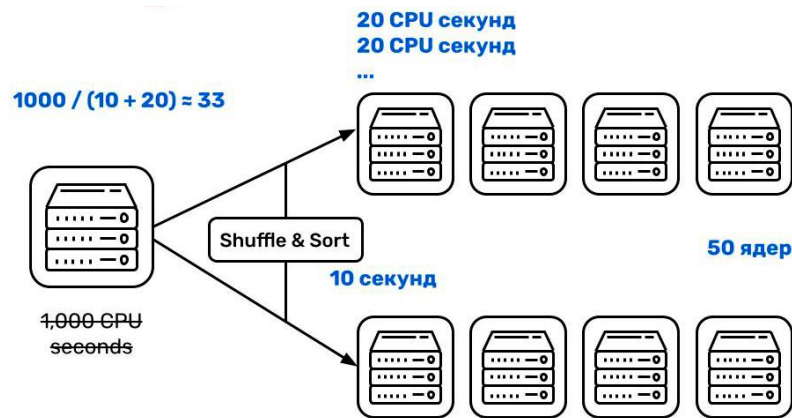
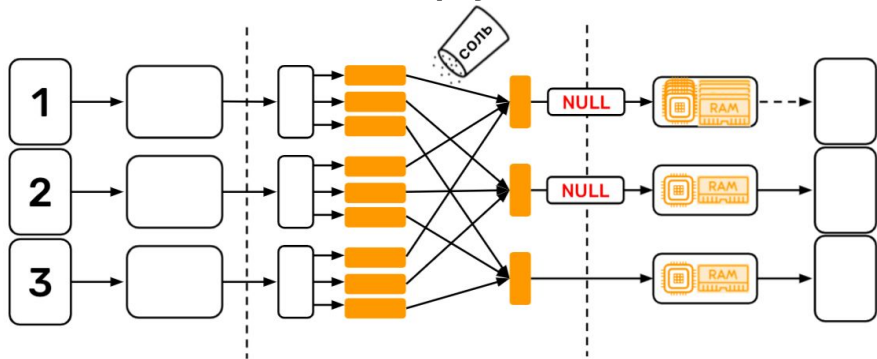
Column-oriented layout (RCFile)

row split 1	row split 2
col1 col2 col3	col1 col2 col3
1 4 2 5 3 6	7 10 8 11 9 12





- ✓ Кодирование vs Сжатие
- ✓ “горячие” vs “холодные” данные
- ✓ Hive: File vs Row format
- ✓ RCFile vs ORC vs Parquet
- ✓ и многое другое





**BIGDATA
TEAM**

А что там в Hadoop 3.0.0?



**BIGDATA
TEAM**

Готовимся к погружению в
математику



Хозяин?
Может не надо...



**BIGDATA
TEAM**

Галуа и умножение



Эварист Галуа́ (фр. Évariste Galois; 1811-1832)

В 18 лет придумал поля* Галуа (1830)

А что сделал ты, когда тебе было 18?

*см. уточнения например [здесь](#) и [здесь](#)



**BIGDATA
TEAM**

Примеры из практики*



Яndex
маркет



Яндекс Новости



Apache
orc™



Parquet

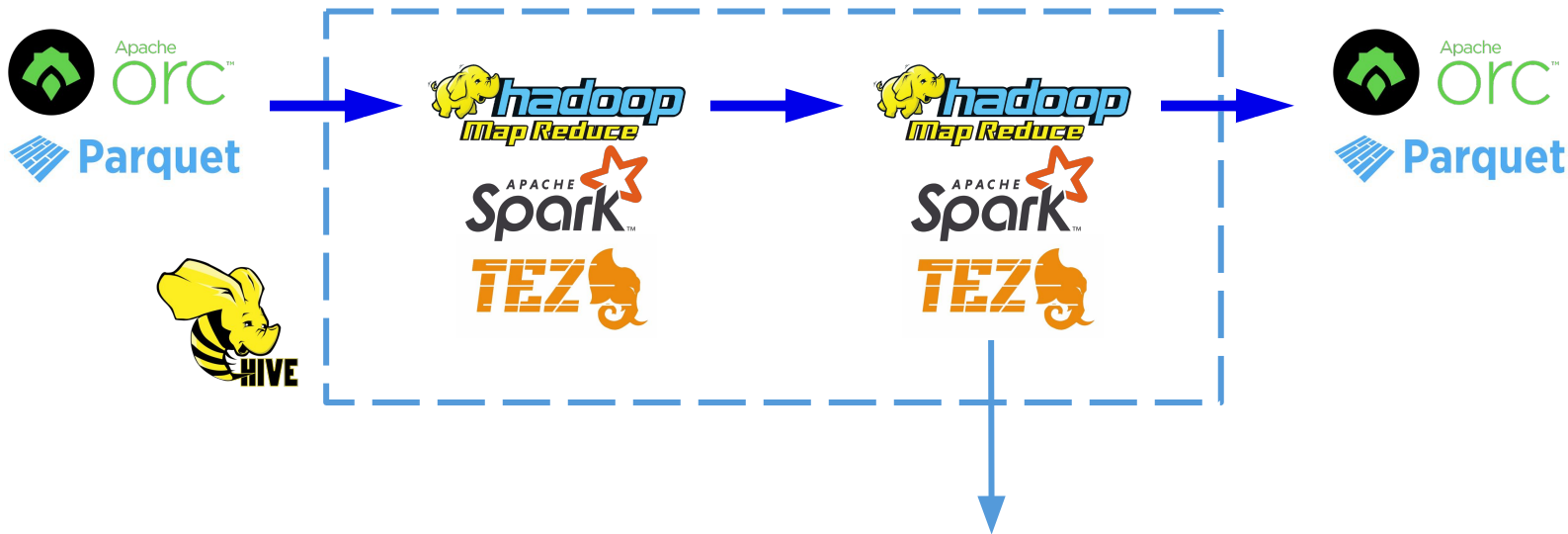


protobuf
Protocol Buffers

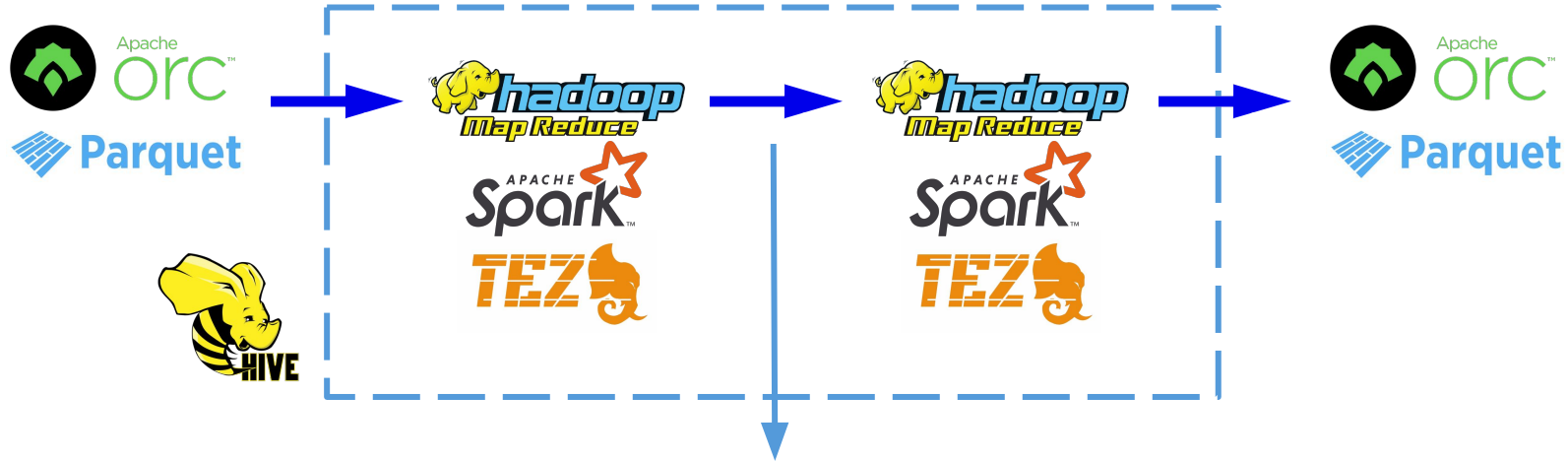
Apache Thrift™



*личный опыт



```
SET mapreduce.map.output.compress=true;  
SET mapreduce.map.output.compress.codec=...;  
SET spark.shuffle.compress=true;  
SET spark.io.compression.codec=...;
```

```
SET hive.exec.compress.intermediate=true;  
SET mapreduce.map.output.compress=true;  
SET mapreduce.map.output.compress.codec=...;  
SET spark.shuffle.compress=true;  
SET spark.io.compression.codec=...;
```