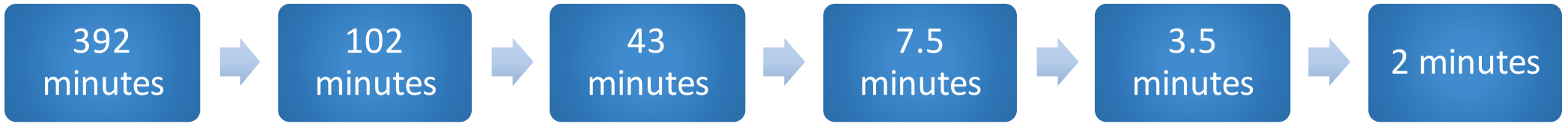
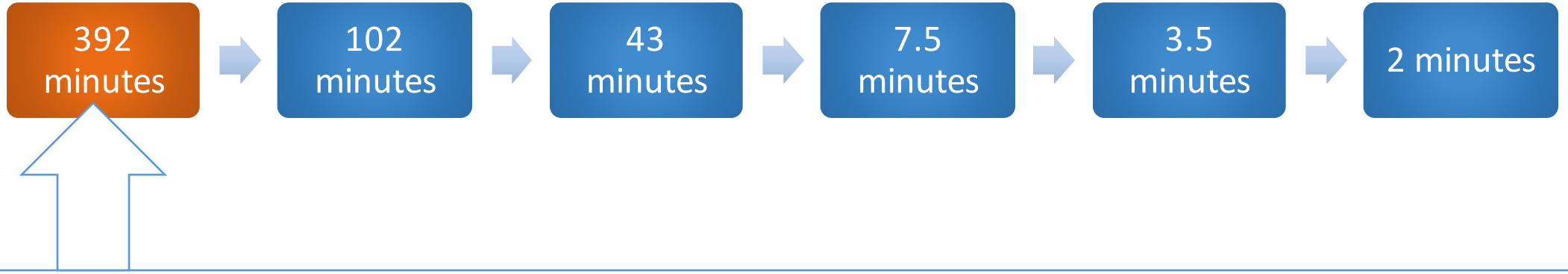
Pocket Data

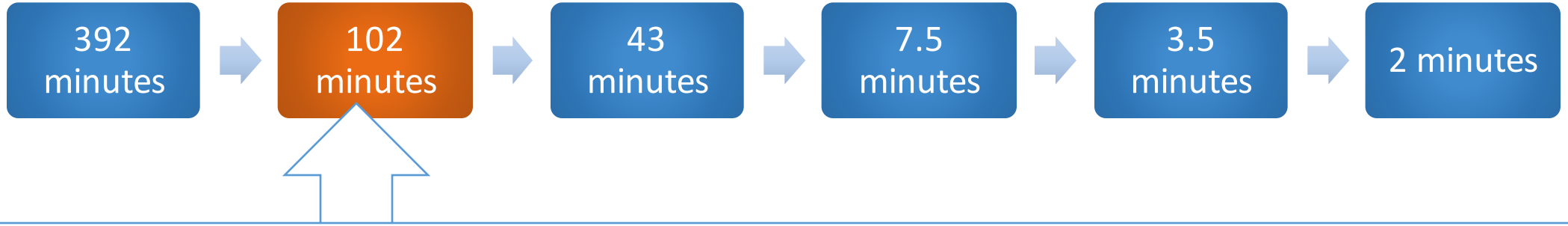
Team : TBD
Naveen, Sankar, Saravanan, Sathish

Parser Evolution

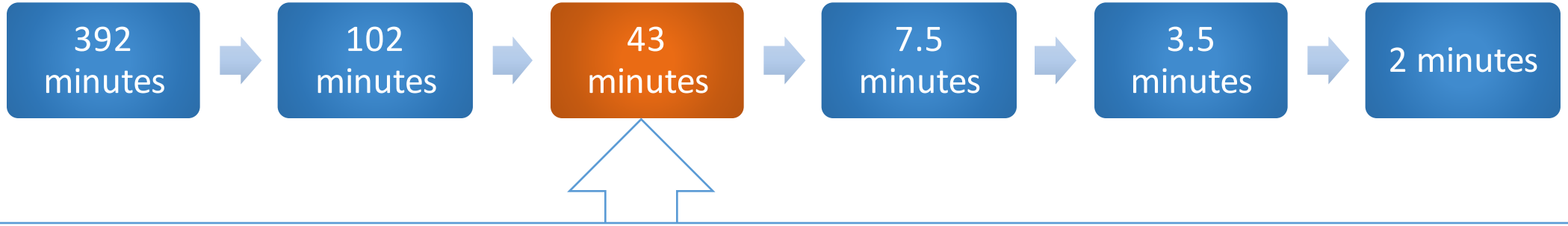




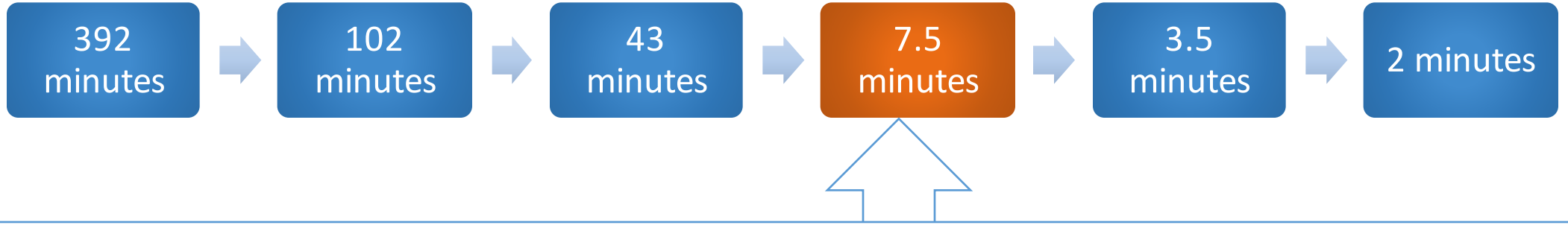
- No serialization
- Sequential Execution
- Parsing logs and Generating analytics took 392.0 minutes.
- Most of the time (over 90%) is taken by jsqlparser to parse SQL and create jsqlparser statement objects.



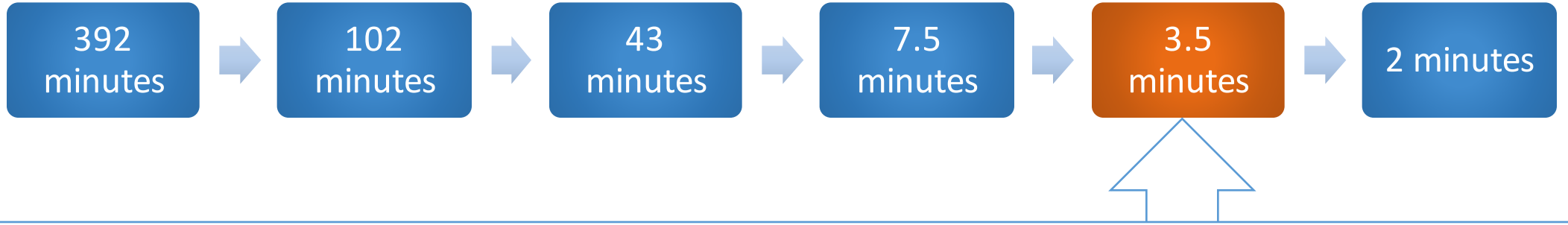
- Jsqlparser bottleneck is solved.
- Split analytics generation into two phases
  - Object serialization – Converting SQL quries to jsqlparser objects, serializing and storing it in file system (took approx 474 minutes)
  - Analytics generation from serialized jsqlparser objects (took 102 minutes) in a sequential manner.
- Object serialization needs to be done only once.



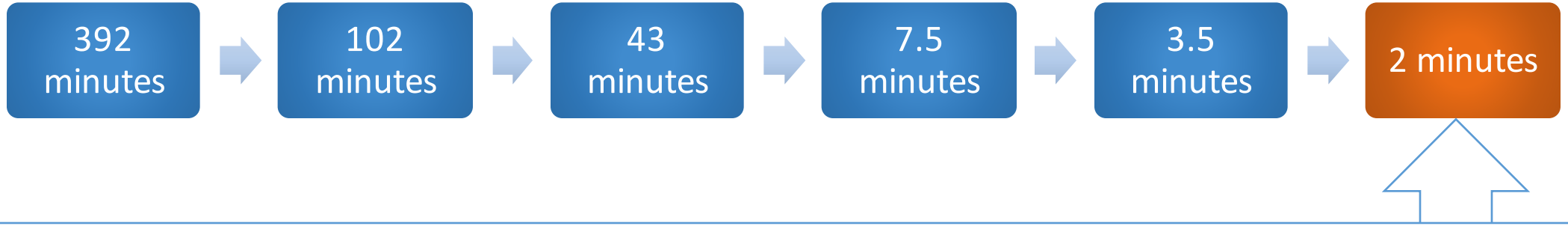
- Reading the objects from file system took 85% of time in analytics generation.
- One thread is assigned for each user's log data
  - 11 threads max
  - Unequal amount of data for each puts more work on few threads
- Improvement: Assigned threads for files instead of user.
  - Configurable number of threads.
  - Got running time to 32 minutes.



- Reading the serialized objects still took 80% of the time.
- Used third party library “Kyro”- specialized for this kind of operation.
- Serialized analytics generation took 7.5 minutes



- Parallel read to generate analytics reduced time to 3.5 times



- Write took more than half of the time.
- Schema generation does not need write like analytics generation.

Schema Generation

- Recreated table schemas from query logs.
- Base version has been implemented without considering constraints.
- Examples
  - SELECT R.a, b FROM R
    - We can infer that a and b are columns of table R
  - SELECT R.a, b FROM R, S
    - In the above query ‘b’ can be in R or S. So, we add them as potential columns in both tables.
  - SELECT a, b FROM R, S where R.a = ‘4’
    - Here, we infer that a belongs to R from the where clause.

```
<table name='wa_contacts' >
    <column name='raw_contact_id' confirmed='true' />
    <column name='photo_ts' confirmed='true' />
    <column name='phone_label' confirmed='true' />
    <column name='phone_type' confirmed='true' />
    <column name='jid' confirmed='true' />
    <column name='sort_name' confirmed='true' />
    <column name='display_name' confirmed='true' />
    <column name='given_name' confirmed='true' />
    <column name='is_whatsapp_user' confirmed='true' />
    <column name='thumb_ts' confirmed='true' />
    <column name='number' confirmed='true' />
    <column name='unseen_msg_count' confirmed='true' />
    <column name='photo_id_timestamp' confirmed='true' />
    <column name='callability' confirmed='true' />
    <column name='wa_name' confirmed='true' />
    <column name='_id' confirmed='true' />
    <column name='status_timestamp' confirmed='true' />
    <column name='family_name' confirmed='true' />
    <column name='status' confirmed='true' />
</table>
```

PRAGMA table_info(wa_contacts)

- Rows returned: 19
- Rows found by schema gen: 19

```
<table name='messages' >
    <column name='media_duration' confirmed='true' />
    <column name='data' confirmed='true' />
    <column name='origin' confirmed='true' />
    <column name='latitude' confirmed='true' />
    <column name='participant_hash' confirmed='true' />
    <column name='media_caption' confirmed='true' />
    <column name='media_url' confirmed='true' />
    <column name='media_hash' confirmed='true' />
    <column name='remote_resource' confirmed='true' />
    <column name='message_table_id' confirmed='false' />
    <column name='media_name' confirmed='true' />
    <column name='raw_data' confirmed='true' />
    <column name='timestamp' confirmed='true' />
    <column name='longitude' confirmed='true' />
    <column name='media_size' confirmed='true' />
    <column name='key_id' confirmed='true' />
    <column name='thumb_image' confirmed='true' />
    <column name='key_from_me' confirmed='true' />
    <column name='recipient_count' confirmed='true' />
    <column name='_id' confirmed='true' />
    <column name='media_mime_type' confirmed='true' />
    <column name='media_wa_type' confirmed='true' />
    <column name='key_remote_jid' confirmed='true' />
    <column name='needs_push' confirmed='true' />
    <column name='status' confirmed='true' />
</table>
```

PRAGMA table_info(messages)

- Cols returned by PRAGMA: 30
- Cols found by schema gen: 25
  - Cols confirmed: 24

```
<table name='chat_list' >
    <column name='media_duration' confirmed='true' />
    <column name='mod_tag' confirmed='true' />
    <column name='data' confirmed='true' />
    <column name='origin' confirmed='true' />
    <column name='latitude' confirmed='true' />
    <column name='participant_hash' confirmed='true' />
    <column name='media_caption' confirmed='true' />
    <column name='sort_timestamp' confirmed='true' />
    <column name='media_url' confirmed='true' />
    <column name='archived' confirmed='true' />
    <column name='media_hash' confirmed='true' />
    <column name='last_read_message_table_id' confirmed='true' />
    <column name='remote_resource' confirmed='true' />
    <column name='last_read_receipt_sent_message_table_id' confirmed='true' />
    <column name='message_table_id' confirmed='true' />
    <column name='media_name' confirmed='true' />
    <column name='raw_data' confirmed='true' />
    <column name='timestamp' confirmed='true' />
    <column name='longitude' confirmed='true' />
    <column name='media_size' confirmed='true' />
    <column name='key_id' confirmed='true' />
    <column name='thumb_image' confirmed='true' />
    <column name='key_from_me' confirmed='true' />
    <column name='recipient_count' confirmed='true' />
    <column name='media_mime_type' confirmed='true' />
    <column name='media_wa_type' confirmed='true' />
    <column name='key_remote_jid' confirmed='true' />
    <column name='needs_push' confirmed='true' />
    <column name='status' confirmed='true' />
</table>
```

PRAGMA table_info(chat_list)

- Rows returned: 10
- Rows found by schema gen: 29

Future Improvements

- SELECT a FROM R, S
  - In the above query ‘a’ can be in R or S. So, we add them as potential columns in both tables.
- SELECT a FROM R where R.a = ‘4’
  - Here, we infer that ‘a’ belongs to R from the where clause.
  - We also infer that column ‘a’ is **NOT** from S
- SELECT a from S,Q
  - We infer that column ‘a’ is from S or Q.
  - We also know that column ‘a’ cannot be from S.
  - So, we infer that column ‘a’ is from Q.