E T L

Talend Fundamentals 2

Cyrus Lentin

Talend - Recap

- Designer
- Palette
- Repository
- Parameters
- Workspace
- Projects
- Metadata
- Jobs

Schema

- Schema is an important inner concept in TOS design;
- Each Row connection must have non-null schema declaration
- Schema defines the dimensions of the data ingoing to and outgoing from a given component;
- Several primitive java types are supported

Schema & Repository

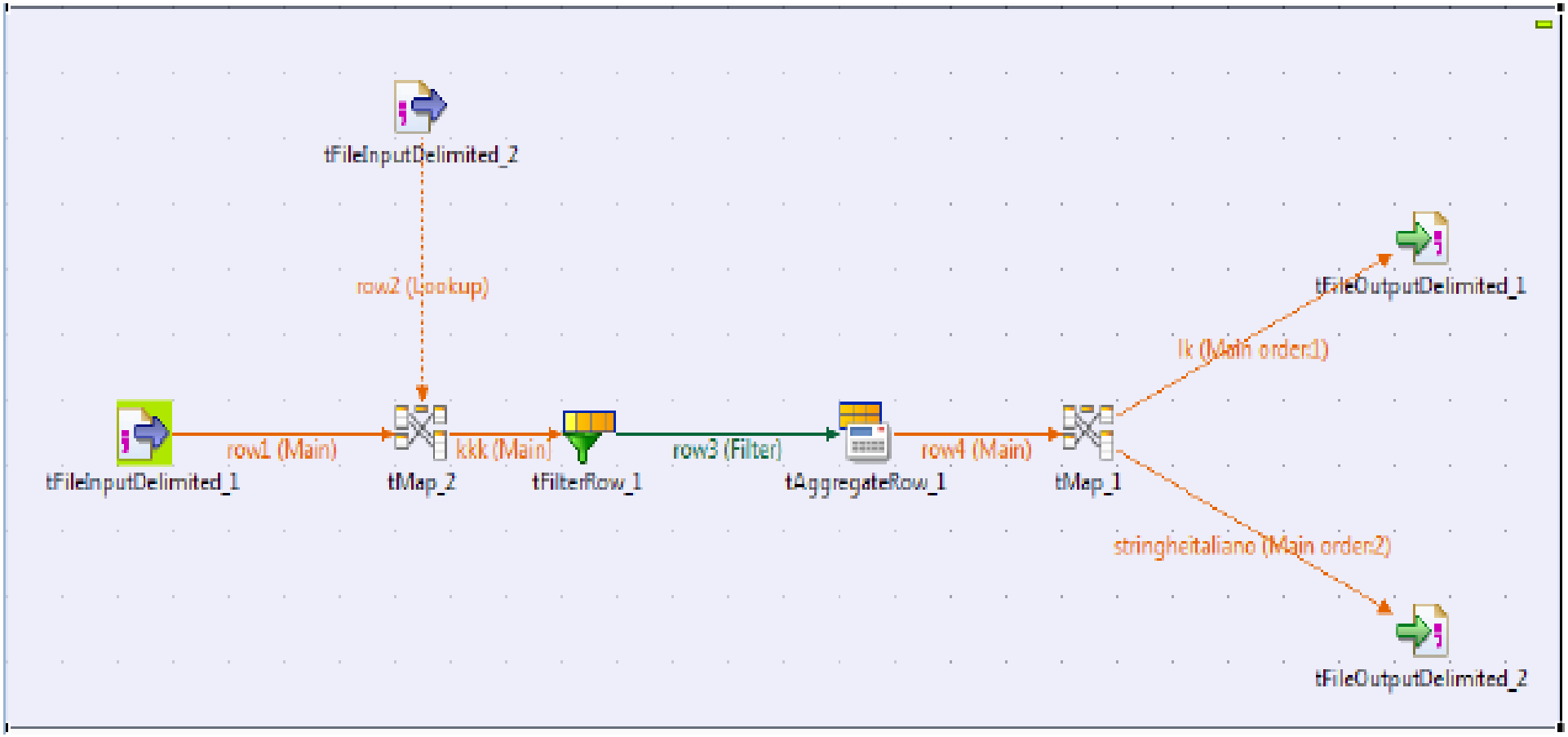
- Very often, Components allows you to select a relevant schema from the Repository;
- Doing so, you will be able to keep parameters between jobs and component instances “in sync”;
- However, this is not mandatory and at any time you can detach the component from the Repository.
- This brings the component in “built in” state, which means that its parameters are locally defined and won’t be updated anymore if the Repository is.

Architecture Of A Job

- A Job is a visual set of components graphically connected using different connections
- From the visual canvas and the connection topology, TOS generates Java code
- This code is procedural by design and not really object oriented
- It's fast
- But debug is a difficult even for the experienced programmer
- In a job, we need to understand the following
  - Sub Jobs
  - Starting Point
  - Ending Point
  - Main Connections
  - Lookup Connections

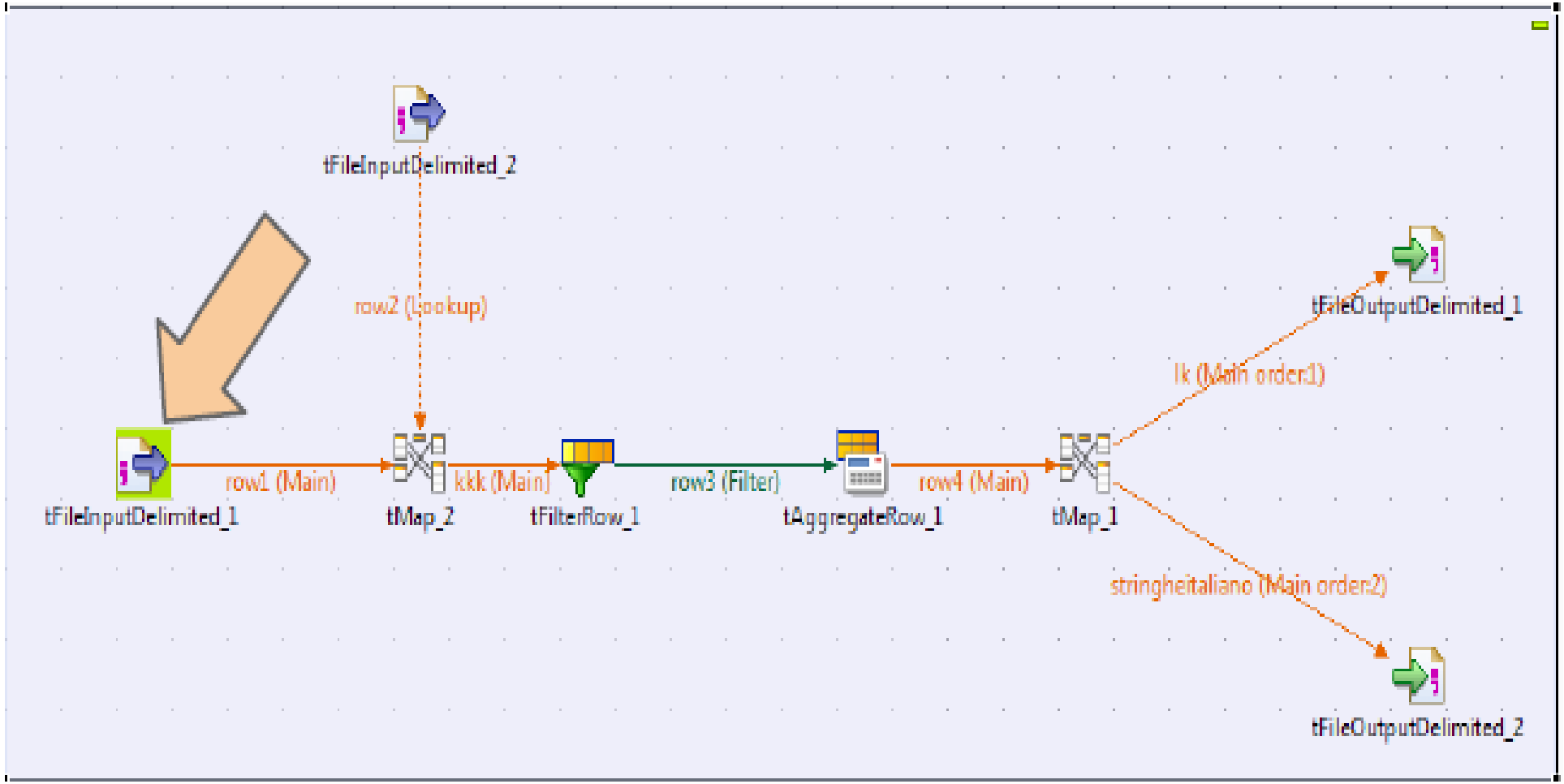
Sub Job

- A set of connected components is part of a subjob if they are all enclosed by a light-blue background
- You can have as many subjobs you need in a given job.



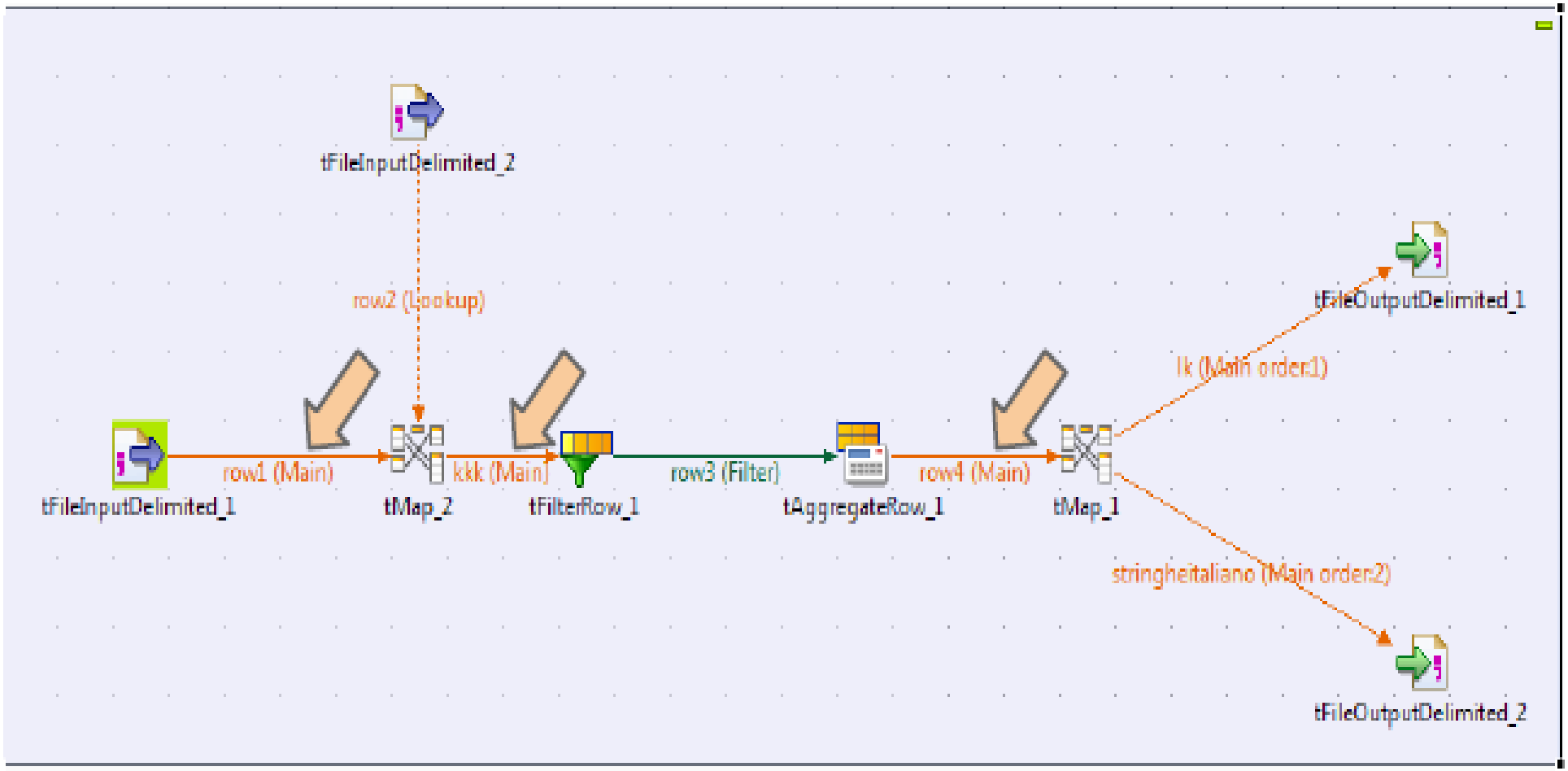
Starting Points

- The starting point component of a subjob is the one with a green background;
- We can have more than one starting points in a subjob, but we cannot decide the execution order!



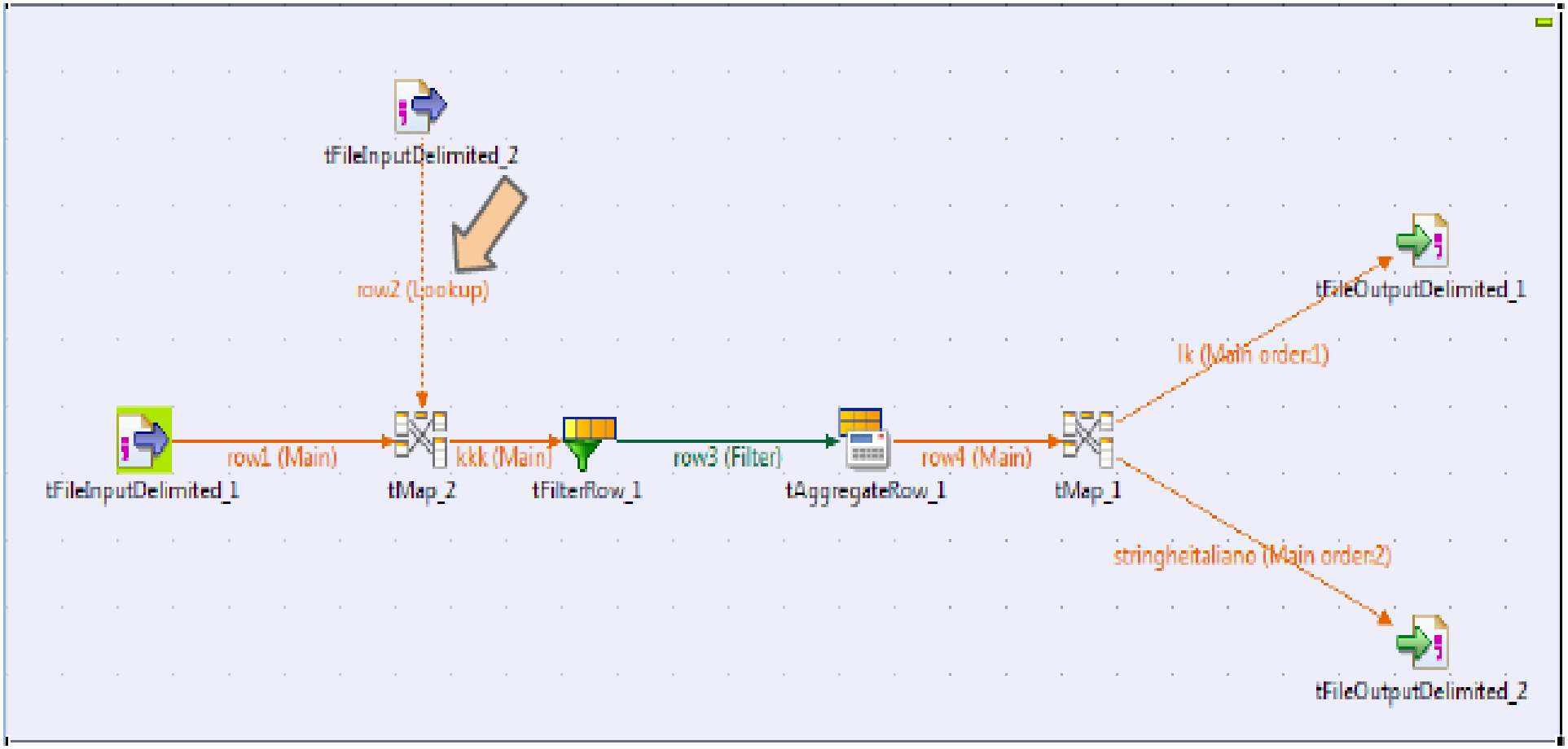
Main Connections

- The Main connections are those that dictate the data flow; they carry data (one row at a time);
- In case you have a split, we can decided the execution priority



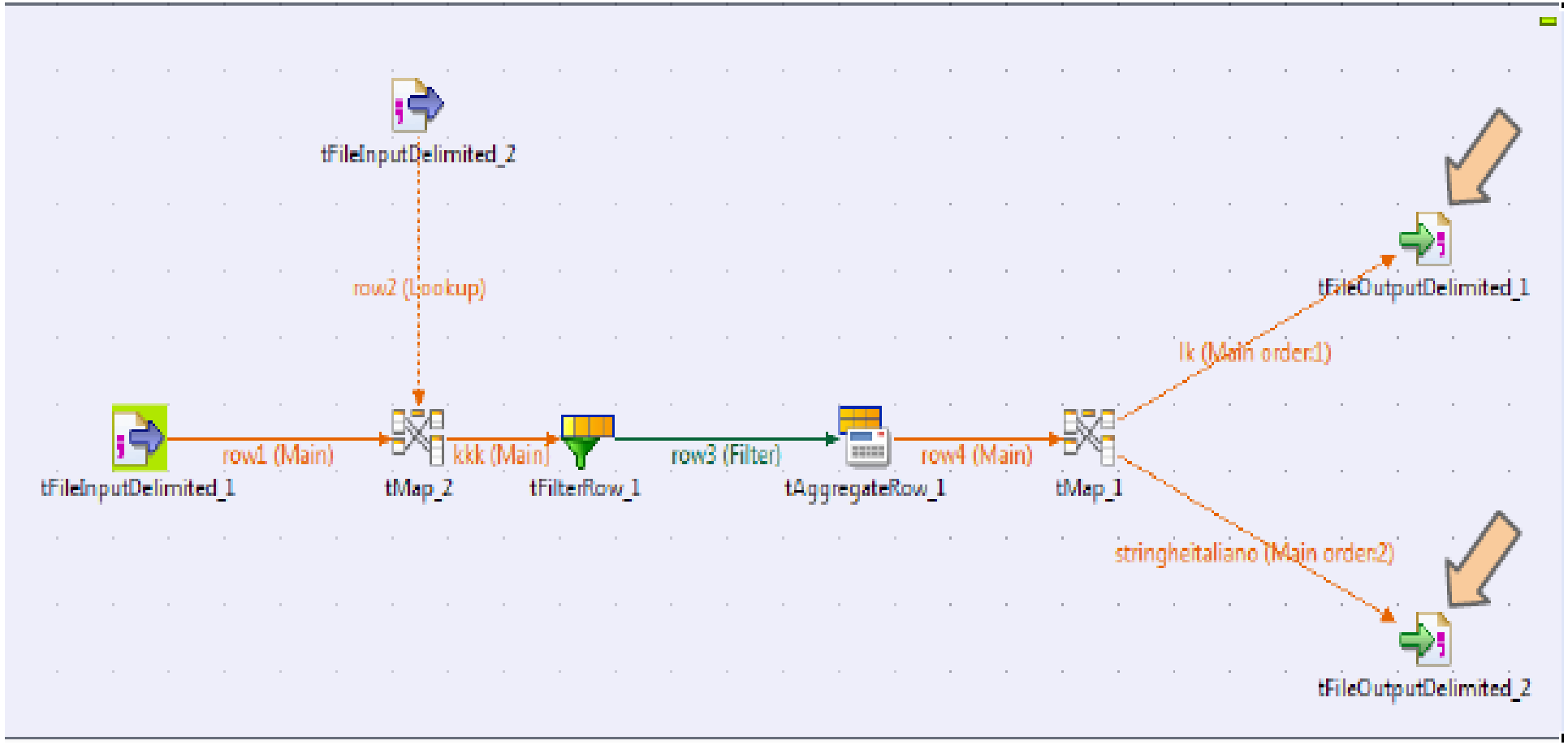
Lookup Connections

- Lookup connections, as the name suggests, make data available for fast-lookup (join or match).
- Typically, lookup data vectors are stored in-memory during job processing; results in OOM issues!



Endpoints

- Endpoints are components that have not outgoing connection.
- A given subjob can have as many endpoints as needed (think about of what's going on after a split)



Connections

There are three types of connections in TOS

- Row
- Trigger
- Iterator

Note: The connections displayed to us are contextual. Not all connections are available for all components. TOS displays only those connections which are allowed for the component.

Row Connection

- Rows are connections that carry on data, one tuple / row at once;
- Their content is defined by a Schema;
- They are used to connect components;
- Components connected this way end up in the same subjob;
- Main, Lookup, Filter, Merge are all components which require data connections;
- Custom components can define their own Data Connection.

Triggers

- Triggers, as the name suggest, won't carry on data, but are actually signals.
- They are usually used to connect subjobs.
- They comes in two main flavours, depending on their scope:
  - Sub Job Triggers
  - Runif Triggers
- They're typically Go/No-Go events to start the execution of one or more subjobs;

Sub Job Triggers

- They are used to connect the starting points of subjobs;
- When connected this way, subjobs will execute sequentially, forcing an execution order;
- You'll end up having only one starting point for the whole chain.

Run If Triggers

- Run If Trigger is a special type of conditional trigger
- Run If Trigger is fired only if embedded expression is evaluated to true
- The expression must be written in Java and have a boolean outcome

Iterators

- Iterators stands in the middle between Data Connections and Triggers;
- They won’t carry on data like Rows…
- ... also they’re not fired only once like Triggers.
- … think of them like Triggers which will be fired once for each incoming row.
- They are connected to starting points, like SubJob Triggers, but originates from standard components like Row Connections.

Parameters

- When you select a component instance, the parameter pane will show the relevant fields show up
- Several types of parameters are allowed: dropdown, radio buttons, schemas, text fields...
- Text fields will often end up writing their value into the generated java code as-is, so be sure to write them properly:
- Enclose strings in double quotes;
- Be sure to match the expected type, or cast otherwise

Thank you!

Contact:

Cyrus Lentin
cyrus@lentins.co.in
+91-98200-94236