PIG

High Level Data Flow Language

Copyright (c) DataFlair Web Services Pvt. Ltd. Restricted circulation

Agenda

- Map-Reduce and the need for Pig Latin
- Pig Latin
- Compilation into Map-Reduce
- Implementation
- Comparison with Map-Reduce
- Optimization in Pig

Copyright (c) DataFlair Web Services Pvt. Ltd. Restricted circulation

The Map-Reduce Appeal

Scale

Scalable due to simpler design

- Only parallelizable operations
- No transactions

$

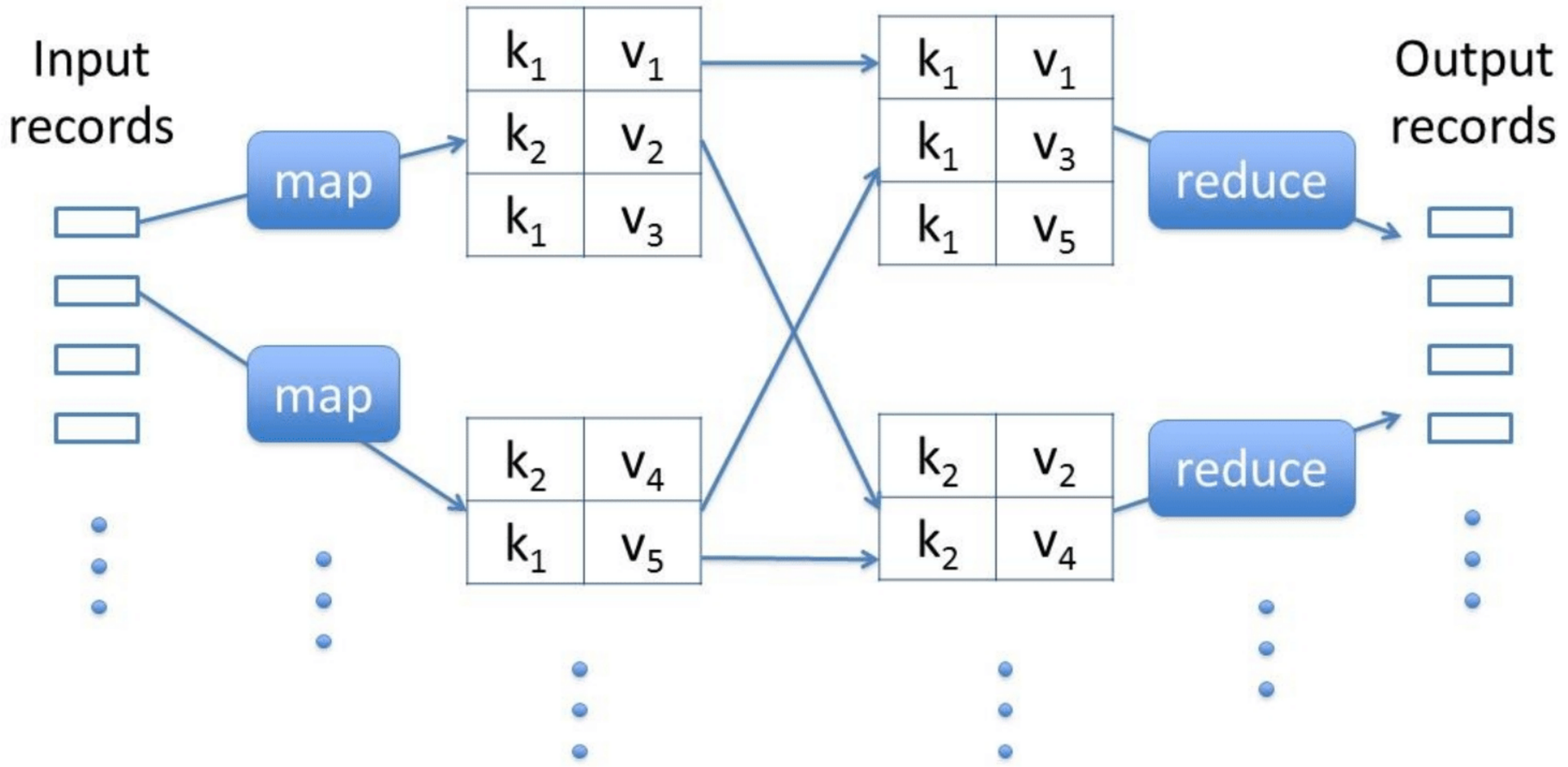
Runs on cheap commodity hardware

~~SQL~~

Procedural Control- a processing “pipe”

Copyright (c) DataFlair Web Services Pvt. Ltd. Restricted circulation

Map-Reduce



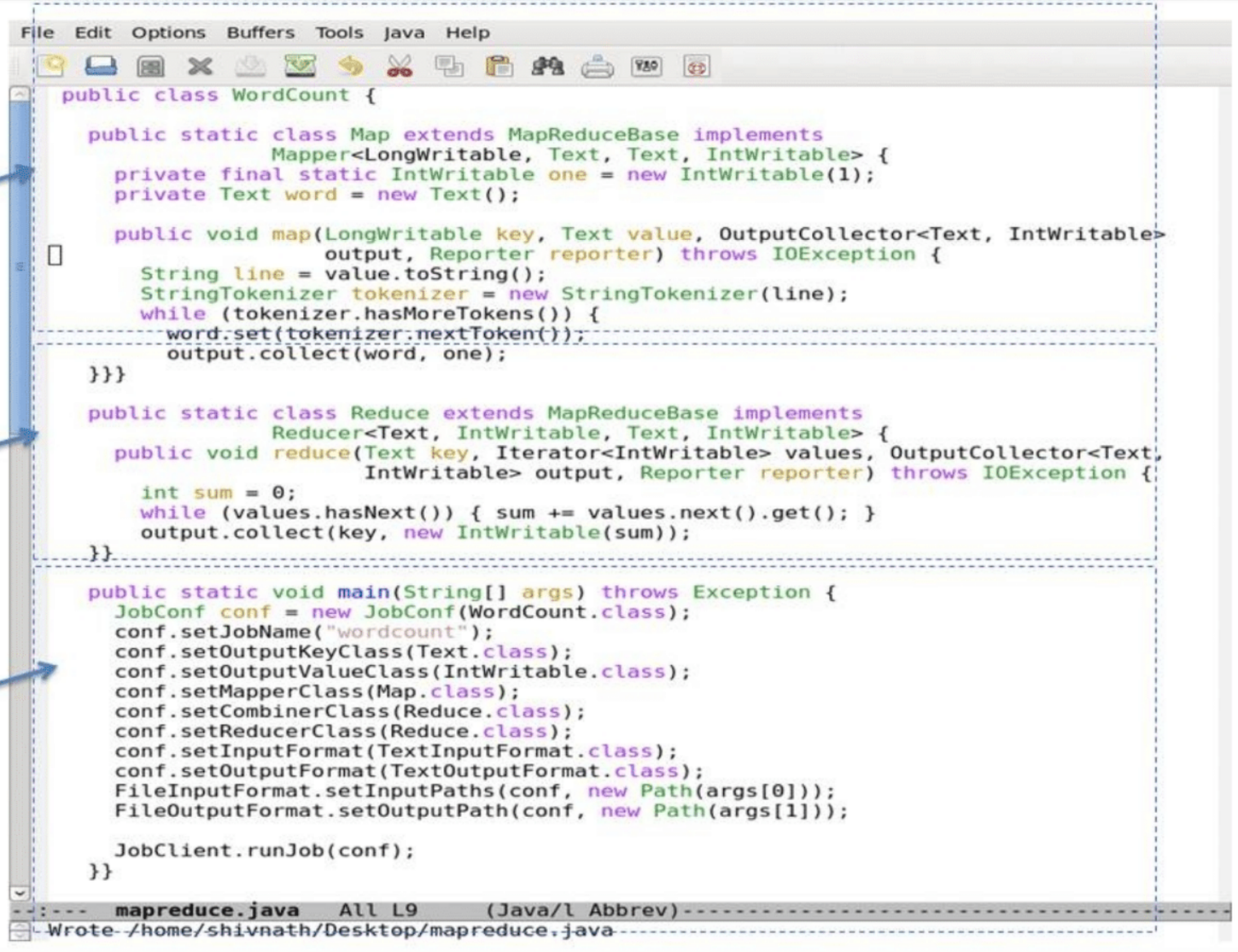
Copyright (c) DataFlair Web Services Pvt. Ltd. Restricted circulation

Java Example

map

reduce

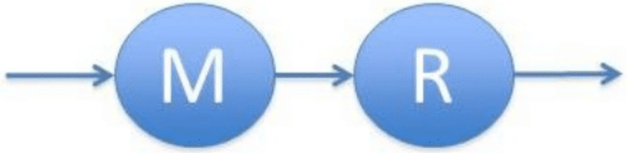
Job conf.



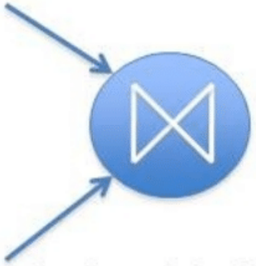
Copyright (c) DataFlair Web Services Pvt. Ltd. Restricted circulation

Disadvantages

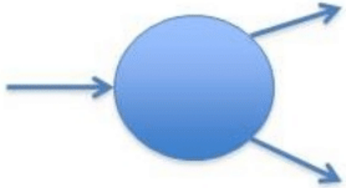
1. Extremely rigid data flow



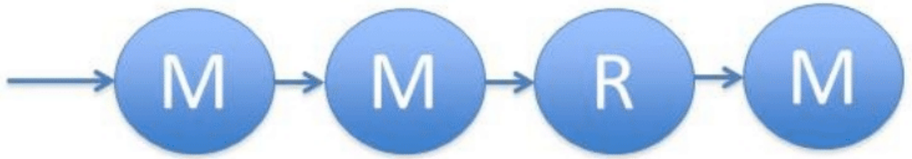
Other flows constantly hacked in



Join, Union



Split



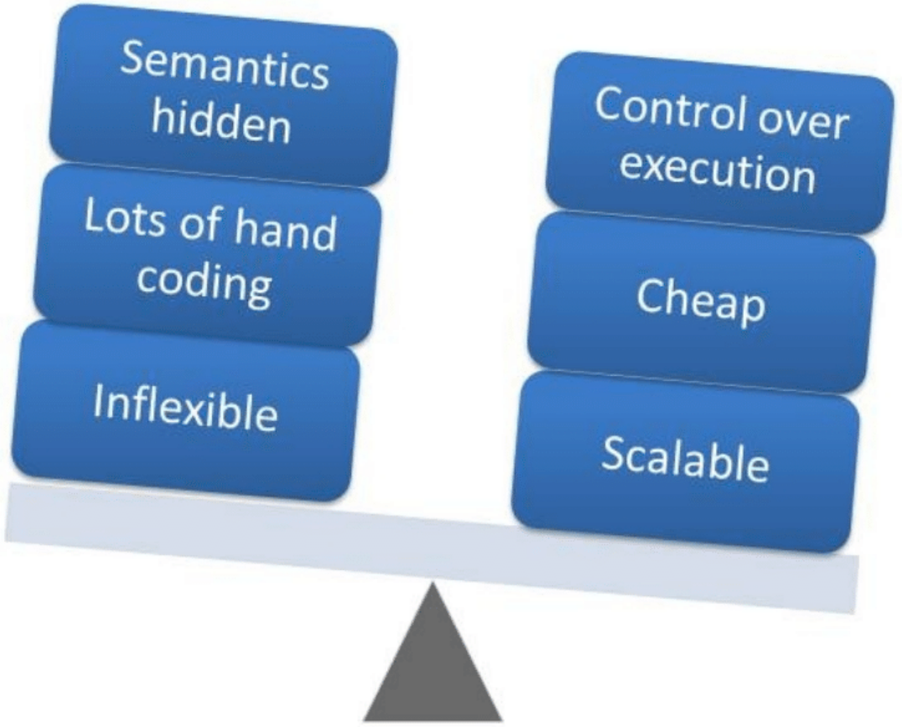
Chains

2. Common operations must be coded by hand
   - Join, filter, projection, aggregates, sorting, distinct

Copyright (c) DataFlair Web Services Pvt. Ltd. Restricted circulation

Pros And Cons

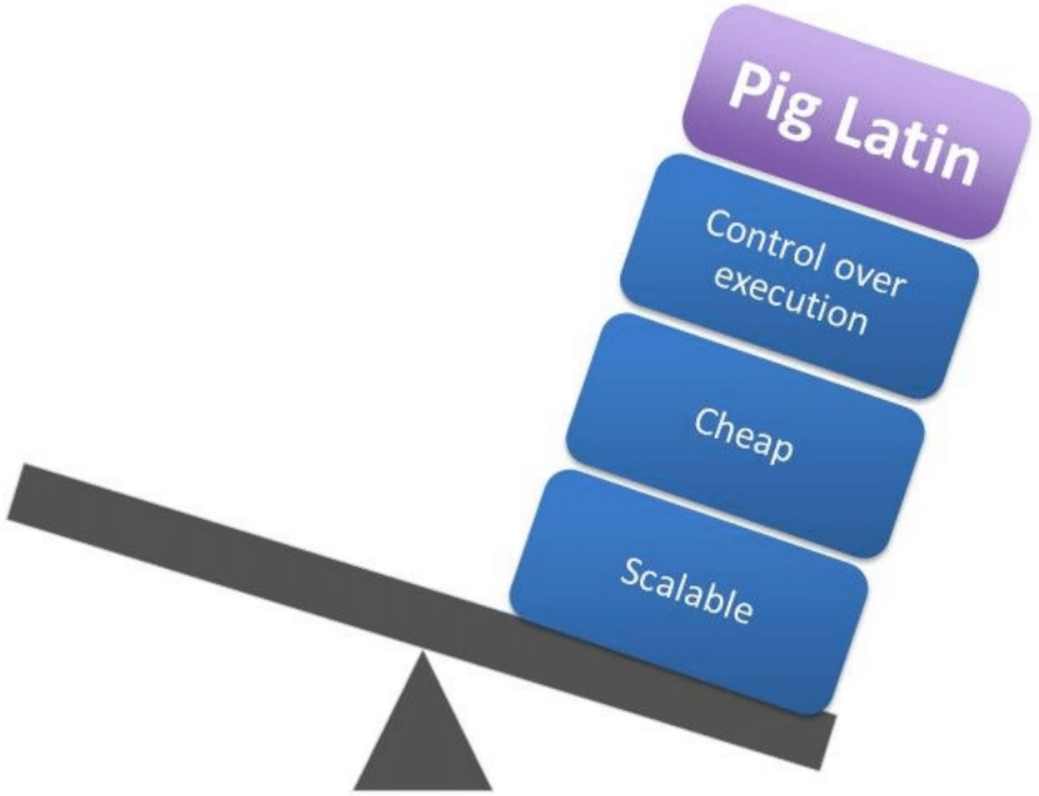
Need a high-level, general data flow language



Copyright (c) DataFlair Web Services Pvt. Ltd. Restricted circulation

Enter Pig Latin

Need a high-level, general data flow language



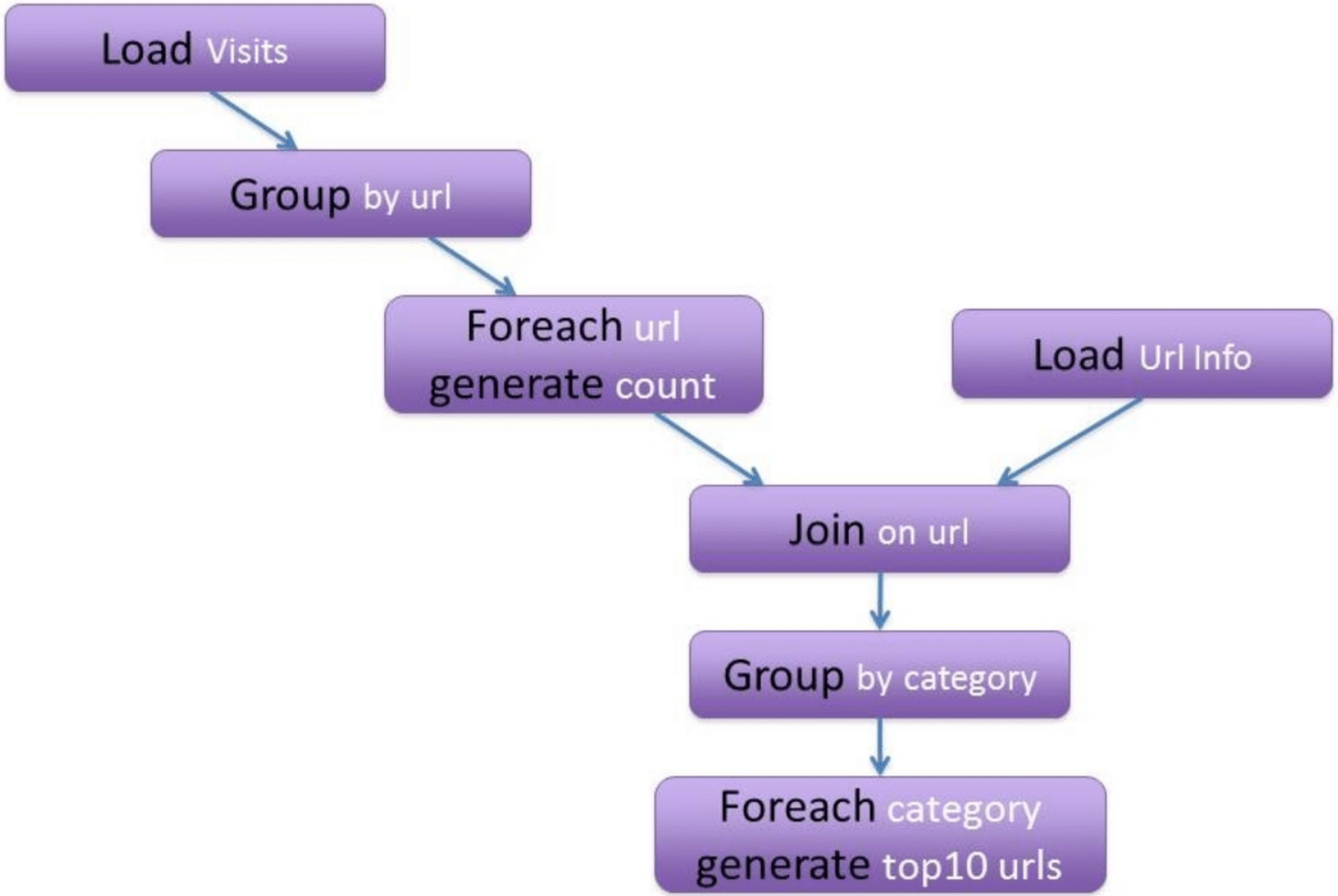
Copyright (c) DataFlair Web Services Pvt. Ltd. Restricted circulation

What is Pig

- A platform for analyzing large data sets that consists of a high-level language for expressing data analysis programs.
- Compiles down to Map Reduce jobs
- Developed by Yahoo!
- Open-source language

Copyright (c) DataFlair Web Services Pvt. Ltd. Restricted circulation

Data Flow



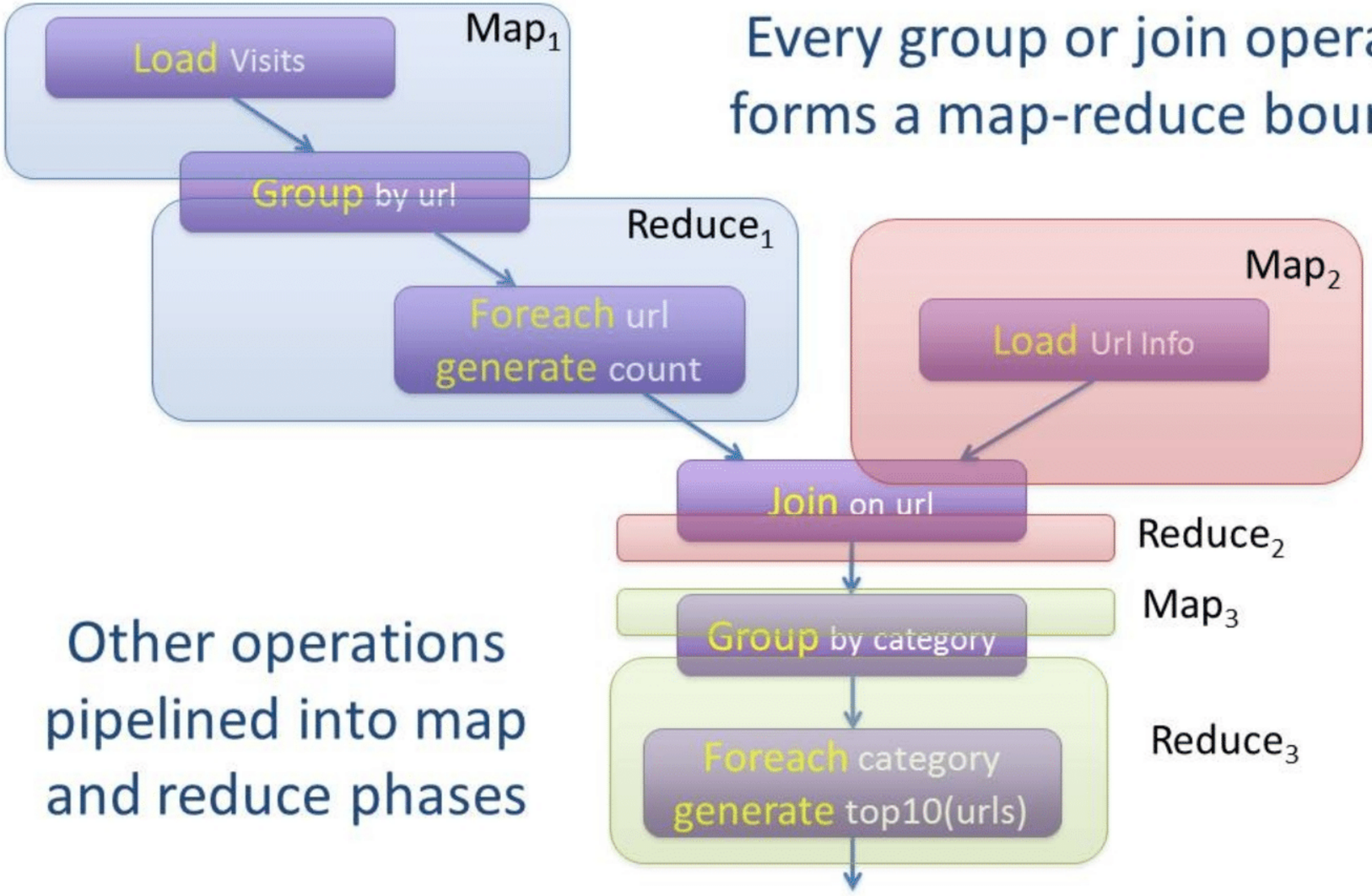
Copyright (c) DataFlair Web Services Pvt. Ltd. Restricted circulation

In Pig Latin

```
visits       = load '/data/visits' as (user, url, time);
gVisits      = group visits by url;
visitCounts  = foreach gVisits generate url, count(visits);

urlInfo      = load '/data/urlInfo' as (url, category, pRank);
visitCounts  = join visitCounts by url, urlInfo by url;

gCategories = group visitCounts by category;
topUrls = foreach gCategories generate top(visitCounts,10);

store topUrls into '/data/topUrls';
```

Copyright (c) DataFlair Web Services Pvt. Ltd. Restricted circulation

Contd..

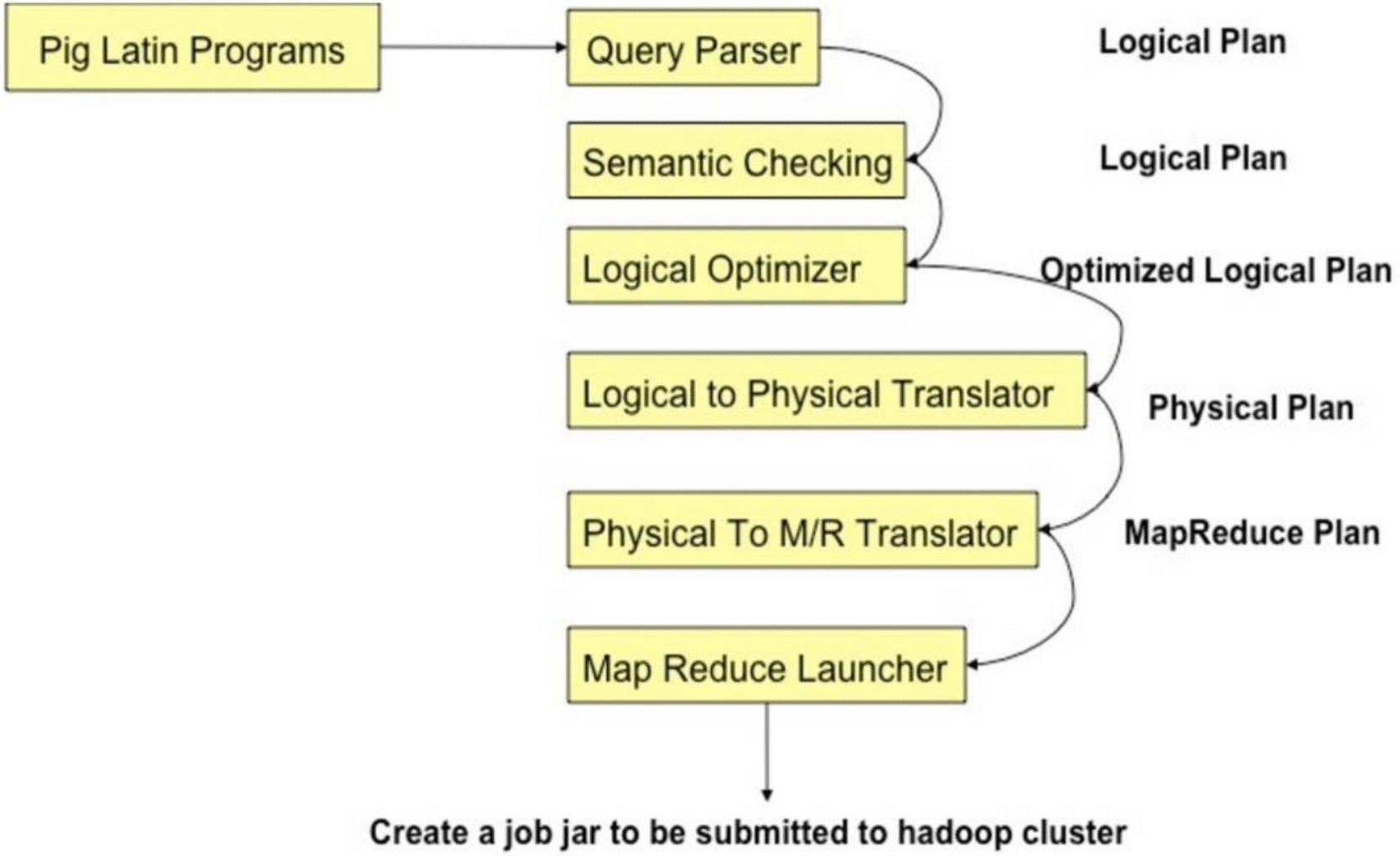


Every group or join operation forms a map-reduce boundary

Other operations pipelined into map and reduce phases

Copyright (c) DataFlair Web Services Pvt. Ltd. Restricted circulation

Pig Compilation



Copyright (c) DataFlair Web Services Pvt. Ltd. Restricted circulation

Implementation