DADSTORM
A simple, fault-tolerant and real-time stream processing system

DAD 2016-2017
Group 12:
Daniel Fermoselle nº 78207
João Marçal nº 78471
Tiago Rodrigues nº 78692

Abstract

DADSTORM is a simple but reliable stream processing system. It's mainly used by Instituto Superior Tecnico Students.
The main features are: the 3 possible semantics of tuple processing it can have, fault-tolerance to f faults per operator with a synchronous model of detection, 3 modes of tuple routing and last but the not the least 5 types of operators. This system is composed by a Puppet Master, Process Creation Service, Operators with their Replicas, Tuples and ThreadPools.

1 Introduction

Nowadays while streaming more and more information is added and we want to process it as fast as possible as well as to get the desired information even though existing the possibility of having faults. In order to get that information in a reliable way we developed DADSTORM. Our system process tuples based on the type of operator which can be UNIQ, COUNT, DUP, FILTER and CUSTOM.

2 Programming Model

In this section, we provide a high-level overview of the programming model, highlighting the key concepts.
In DADSTORM data is represented as string Tuples, i.e, a collection of strings.
To begin with we have a puppet master who will start all the operators on all the machines with the help of the process creation service this last will already be located in all the machines that will run operator replicas.
The puppet master has an intuitive interface inside which we have a box where we can introduce a path to a configuration file to start all the operator replicas as well has their inputs, routing, operator spec and address.
The tuples are stored inside a file that will be accessed by the first operator of the stream. The operator are composed by repicas which can be located inside different machines.
All the replicas might fail but we assure f fault tolerance to silent failures but we assume that none of the first operator replicas can fail and that at least there is one replica alive per operator to guarantee that we have at least f+1 replicas.

3 DADSTORM Abstractions

3.1 Tuple

3.2 Operators and Replicas

3.3 RepInfo

3.4 ConfigInfo

4 Architecture and Implementation

4.1 Tuple

4.2 Operators and Replicas

4.3 Puppet Master

4.4 Process Creation Service

5 Discussion

6 Production Experiences

7 Evaluation

8 Conclusion