

Practical Work with a Gaia Mock Catalogue

Daniel Tapiador

ESA/ESAC

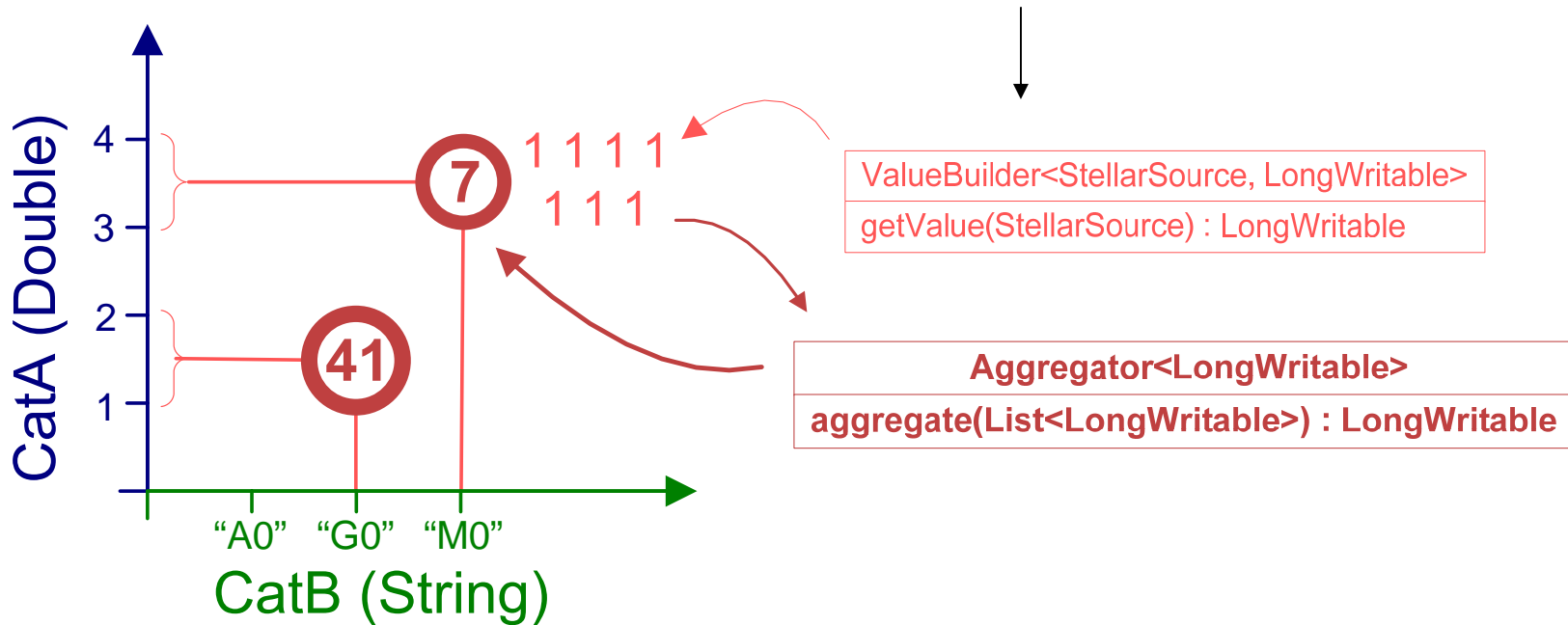
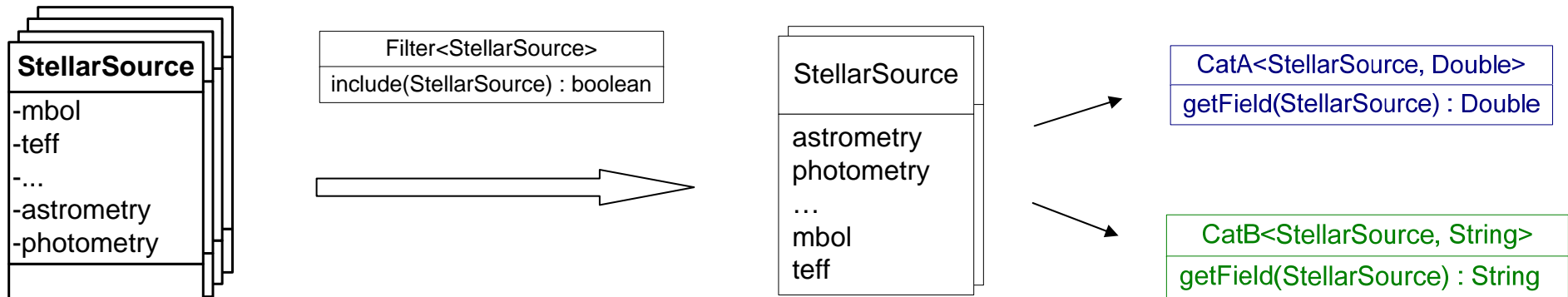
1st March 2012

- Databases...
 - Give a good performance up to a determined amount of data.
 - Ideal for **searching** data with a query language (SQL).
 - We'll definitely provide this type of interface (**TAP**).

- MapReduce is a new data processing paradigm for dealing with big data.
 - Not just a search of information. Some **computation involved** as well.
 - Originated at **Google**. Big data companies making use of this extensively (Google, Facebook, Twitter, LinkedIn, Amazon, Yahoo! Etc).
 - Ideal for **data mining** (for discovery), machine learning (predictive analysis, complex patterns recognition), etc.

- Pre-computed statistics generation is normally a **time consuming** task which must be performed for **every release** of data (validation, summary, etc).
- It involves the generation of **n-dimensional histograms**.
- Create as many histograms/statistics/plots as possible in a **single scan** of data.
- MapReduce is the **natural way** to do this:
 - Brute force always needed.
 - A good scalability and speedup is a **MUST**.

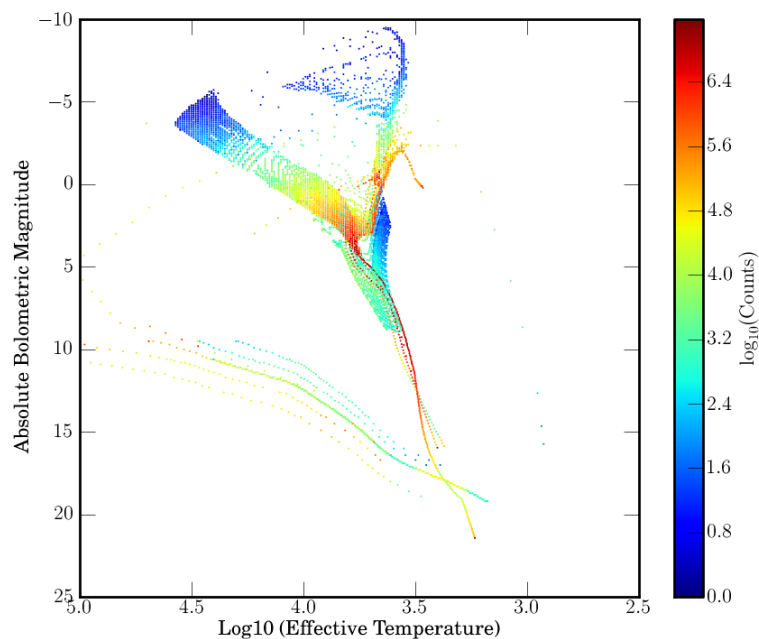
Framework Data Flow



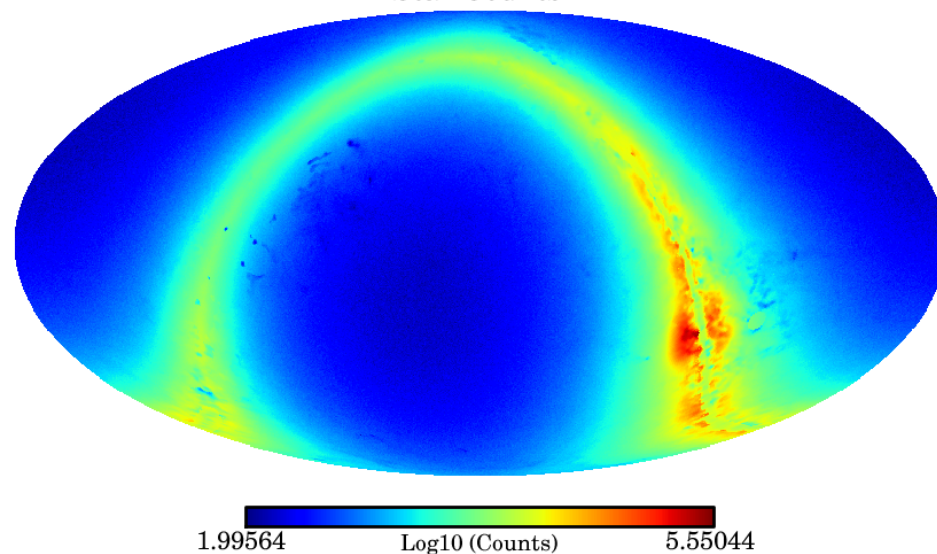
Examples (GUMS-10)

- Theoretical HR
- 1h15min (16 worker nodes)

Theoretical Hertzsprung-Russell Diagram



Star Counts

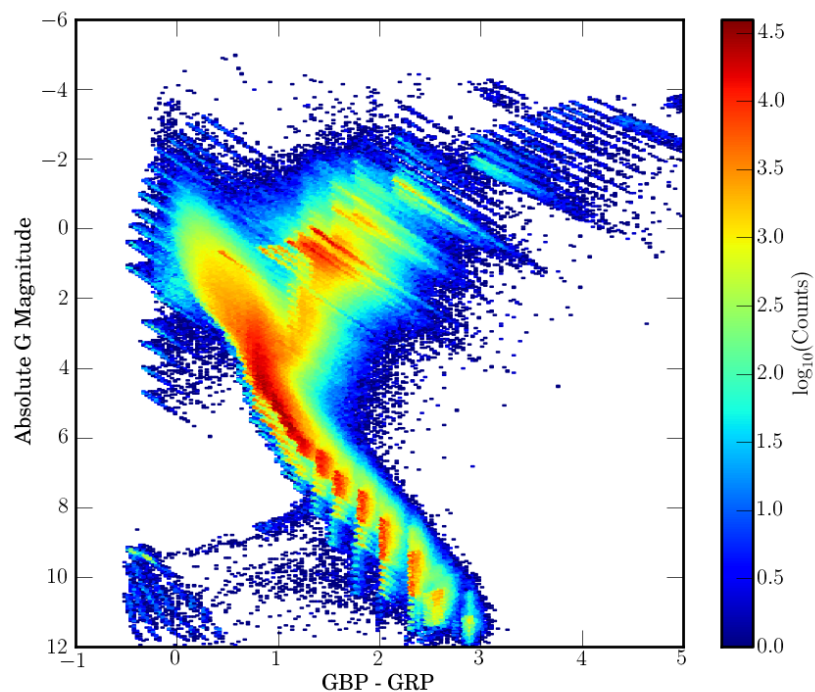


- Healpix Map (NSIDE 256)
- 2h15m (8 worker nodes)

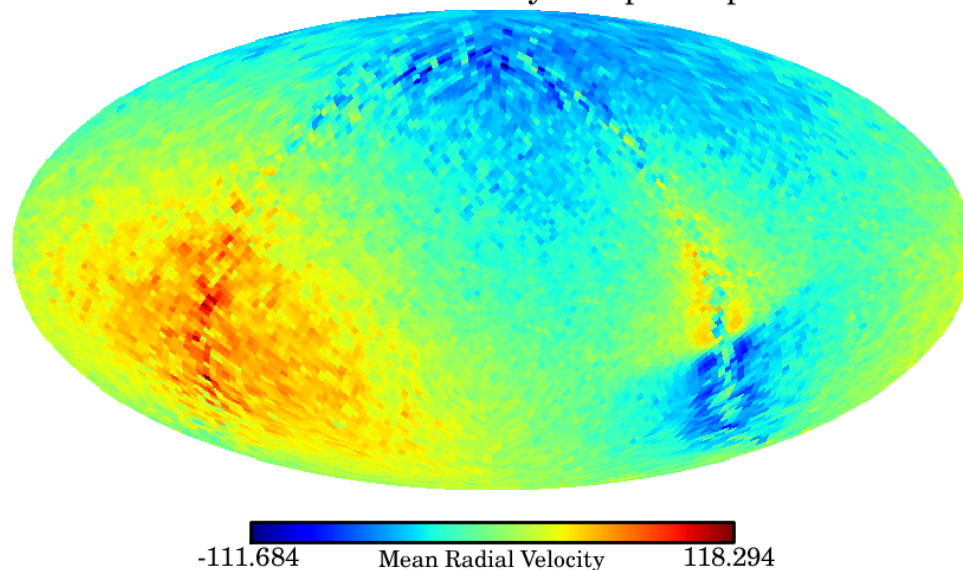
Examples (GOG-17)

- Colour-Magnitude Histogram
- 1h25min (16 worker nodes)

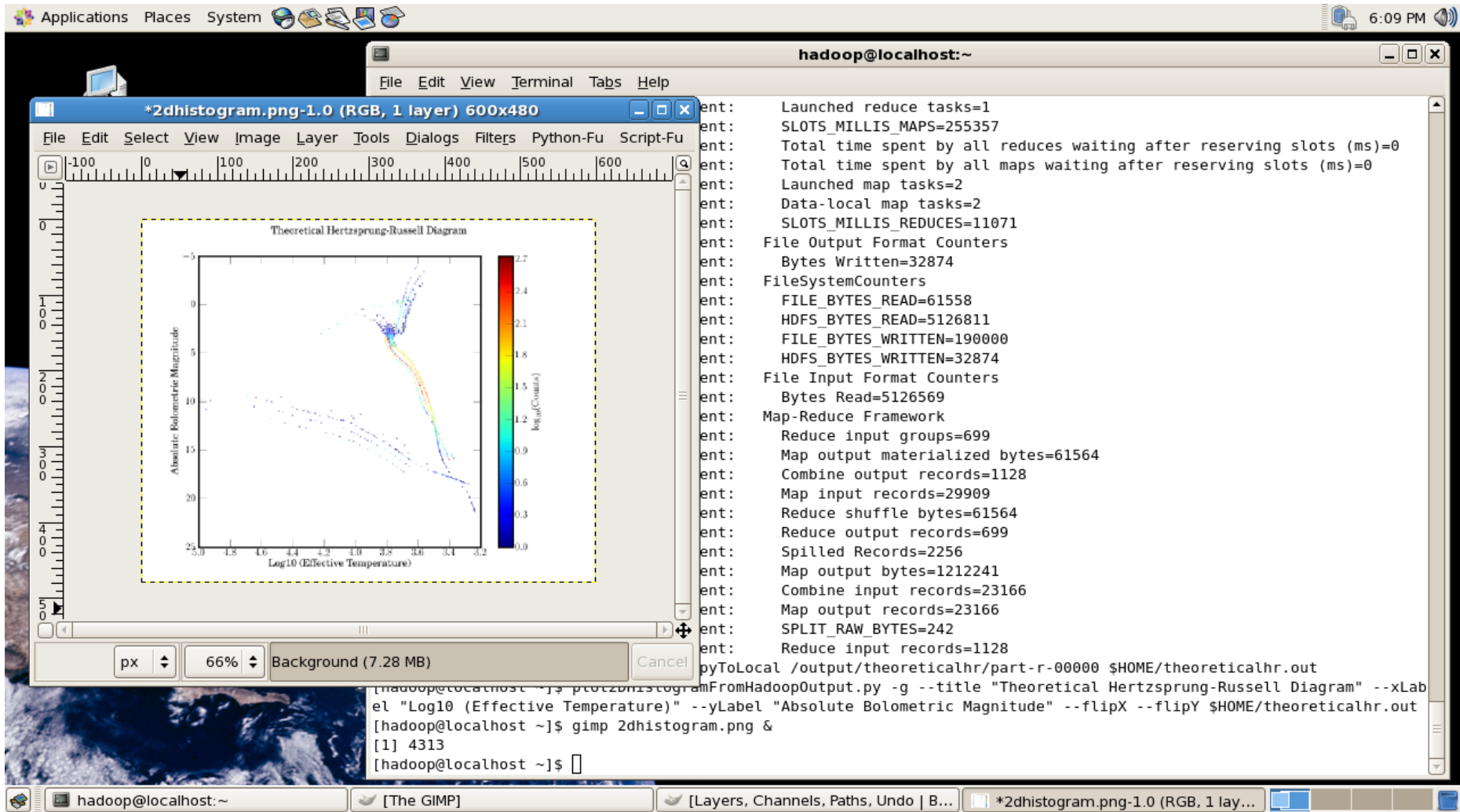
Colour-Magnitude Diagram



Mean Radial Velocity Healpix Map



- Mean Radial Velocity Map (NSIDE 32)
- 1h20min (16 worker nodes)



The screenshot shows a Linux desktop environment with a terminal window and a GIMP window. The terminal window displays the output of a Hadoop job, including metrics for map and reduce tasks, file I/O, and network shuffle. The GIMP window shows a Hertzsprung-Russell diagram titled "Theoretical Hertzsprung-Russell Diagram". The x-axis is labeled "Log10 (Effective Temperature)" and ranges from 2.0 to 5.0. The y-axis is labeled "Absolute Bolometric Magnitude" and ranges from -5 to 25. The plot shows a distribution of stars with a color bar on the right indicating the number of stars per bin, ranging from 0.0 to 2.7.

```
hadoop@localhost:~$ cat /dev/null > /tmp/hadoop-logs.txt
hadoop@localhost:~$ hadoop jar /usr/share/hadoop-mapreduce/hadoop-mapreduce-examples.jar wordcount /tmp/hadoop-logs.txt
Launched reduce tasks=1
SLOTS_MILLIS_MAPS=255357
Total time spent by all reduces waiting after reserving slots (ms)=0
Total time spent by all maps waiting after reserving slots (ms)=0
Launched map tasks=2
Data-local map tasks=2
SLOTS_MILLIS_REDUCE=11071
File Output Format Counters
  Bytes Written=32874
FileSystemCounters
  FILE_BYTES_READ=61558
  HDFS_BYTES_READ=5126811
  FILE_BYTES_WRITTEN=190000
  HDFS_BYTES_WRITTEN=32874
File Input Format Counters
  Bytes Read=5126569
Map-Reduce Framework
  Reduce input groups=699
  Map output materialized bytes=61564
  Combine output records=1128
  Map input records=29909
  Reduce shuffle bytes=61564
  Reduce output records=699
  Spilled Records=2256
  Map output bytes=1212241
  Combine input records=23166
  Map output records=23166
  SPLIT_RAW_BYTES=242
  Reduce input records=1128
pyToLocal /output/theoreticalhr/part-r-00000 $HOME/theoreticalhr.out
[hadoop@localhost ~]$ python2.7 plot2dhistogramFromHadoopOutput.py -g --title "Theoretical Hertzsprung-Russell Diagram" --xLabel "Log10 (Effective Temperature)" --yLabel "Absolute Bolometric Magnitude" --flipX --flipY $HOME/theoreticalhr.out
[hadoop@localhost ~]$ gimp 2dhistogram.png &
[1] 4313
[hadoop@localhost ~]$
```