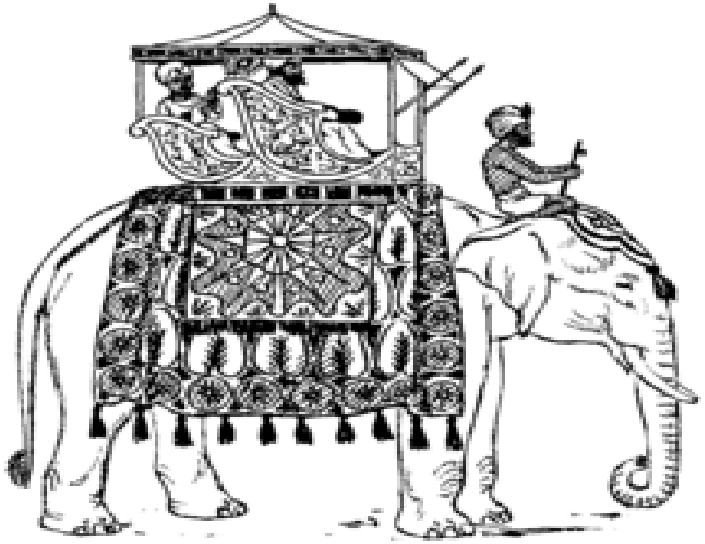
Howdah

a flexible pipeline framework and applications
to analyzing genomic data



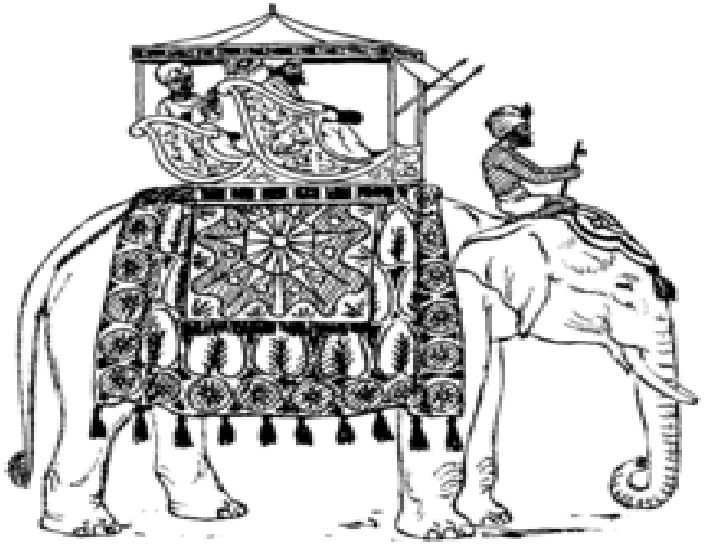
Steven Lewis PhD
slewis@systemsbiology.org

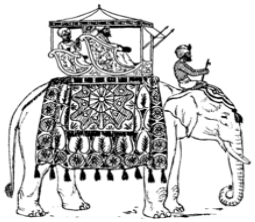
INSTITUTE FOR Systems Biology™

Revolutionizing science. Enhancing life.

What is a Howdah?

- A howdah is a carrier for an elephant
- The idea is that multiple tasks can be performed during a single Map Reduce pass





INSTITUTE FOR Systems Biology
Revolutionizing science. Enhancing life.

Why Howdah?

- Many of the jobs we perform in biology are structured
  - The structure of the data is well known
  - The operations are well known
  - The structure of the output is well known and simple concatenation will not work.
- We need to perform multiple operations with multiple output files on a single data set

INSTITUTE FOR Systems Biology
Revolutionizing science. Enhancing life.

Why Howdah?

- Many of the jobs we perform in biology are structured
  - The structure of the data is well known
  - The operations are well known
  - The structure of the output is well known and simple concatenation will not work.
- We need to perform multiple operations with multiple output files on a single data set

INSTITUTE FOR Systems Biology
Revolutionizing science. Enhancing life.

General Assumptions

- The data set being processed is large and a Hadoop map-reduce step is relatively expensive
- The final output set is much smaller than the input data and is not expensive to process
- Further steps in the processing may not be handled by the cluster
- Output files require specific structure and formatting

INSTITUTE FOR Systems Biology
Revolutionizing science. Enhancing life.

Why Not Cascade or Pig

- Much or the code in biological processing is custom
- Special Formats
- Frequent exits to external code such as python
- Output must be formatted and usually outside of HDFS

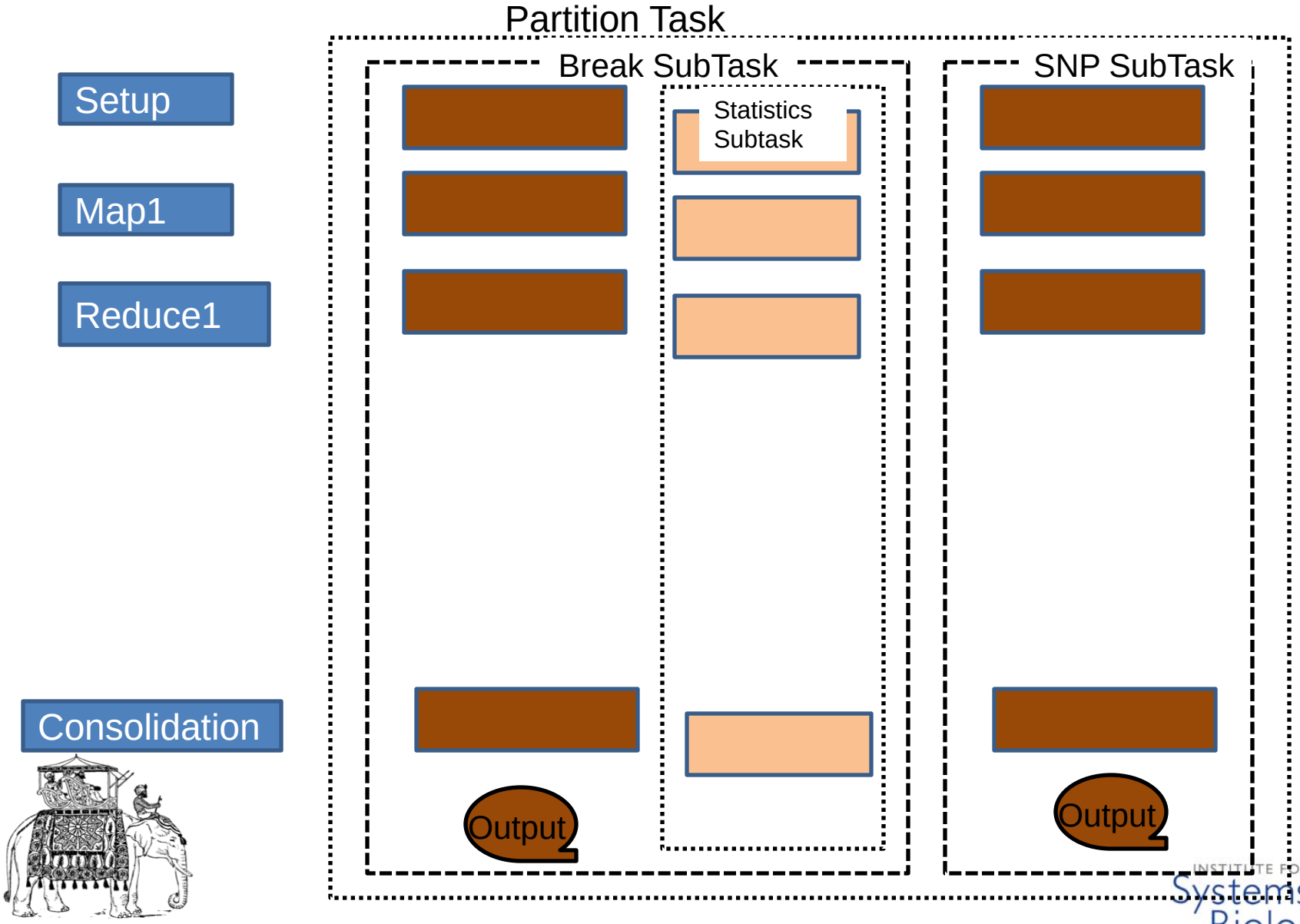
INSTITUTE FOR Systems Biology
Revolutionizing science. Enhancing life.

Job -> Multiple Howdah Tasks

- Howdah tasks pick up data during a set of Map-Reduce jobs
- Task own their data prepending markers to the keys
- Task may spawn multiple sub-tasks
- Tasks (and subtasks) may manage their ultimate output
- Howdah tasks exist at every phase of a job
  including pre and post launch

INSTITUTE FOR Systems Biology
Revolutionizing science. Enhancing life.

Howdah Tasks



INSTITUTE FOR Systems Biology
Revolutionizing science. Enhancing life.

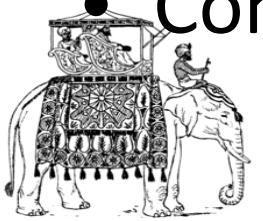
Task Life Cycle

- Tasks are created by reading an array of Strings in the job config.
- The strings creates instances of Java classes and sets parameters
- All tasks are created before the main job is run to allow each task to add configuration data.
- Tasks are created in all steps of the job but are often inactive.

INSTITUTE FOR Systems Biology™
Revolutionizing science. Enhancing life.

Howdah Phases

- Job Setup – before the job starts – sets up input files, configuration, distributed cache
- Processing
  - Initial Map – data incoming from files
  - Map(n) Subsequent maps – data assigned to a task
  - Reduce(n) – data assigned to a task
- Consolidation – data assigned to a path



INSTITUTE FOR Systems Biology
Revolutionizing science. Enhancing life.

Critical Concepts

Looking at the kinds of problems we were solving we found several common themes

- Multiple action streams in a single process
- Broadcast and Early Computation
- Consolidation

INSTITUTE FOR Systems Biology™
Revolutionizing science. Enhancing life.

Broadcast

- The basic problem
  - Consider the case where all reviewers need access to a number of global totals.
    - Sample - a Word Count program wants to not only output the count but the fraction of all words of a specific length represented by this word. Thus the word "a" might by 67% of all words of length 1.
    - Real – a system is interested in the probability that a reading will be seen. Once millions of readings have been observed, probability is the fraction of readings whose values are >= the test reading.

Maintaining Statistics

- For all such cases the processing needs access to global data and totals
- Consider the problem of counting the number of words of a specific length.
  - It is trivial for every mapper to keep count of the number of words of a particular length observed.
  - It is also trivial to send this data as a key/value pair in the cleanup operation.

INSTITUTE FOR Systems Biology™
Revolutionizing science. Enhancing life.

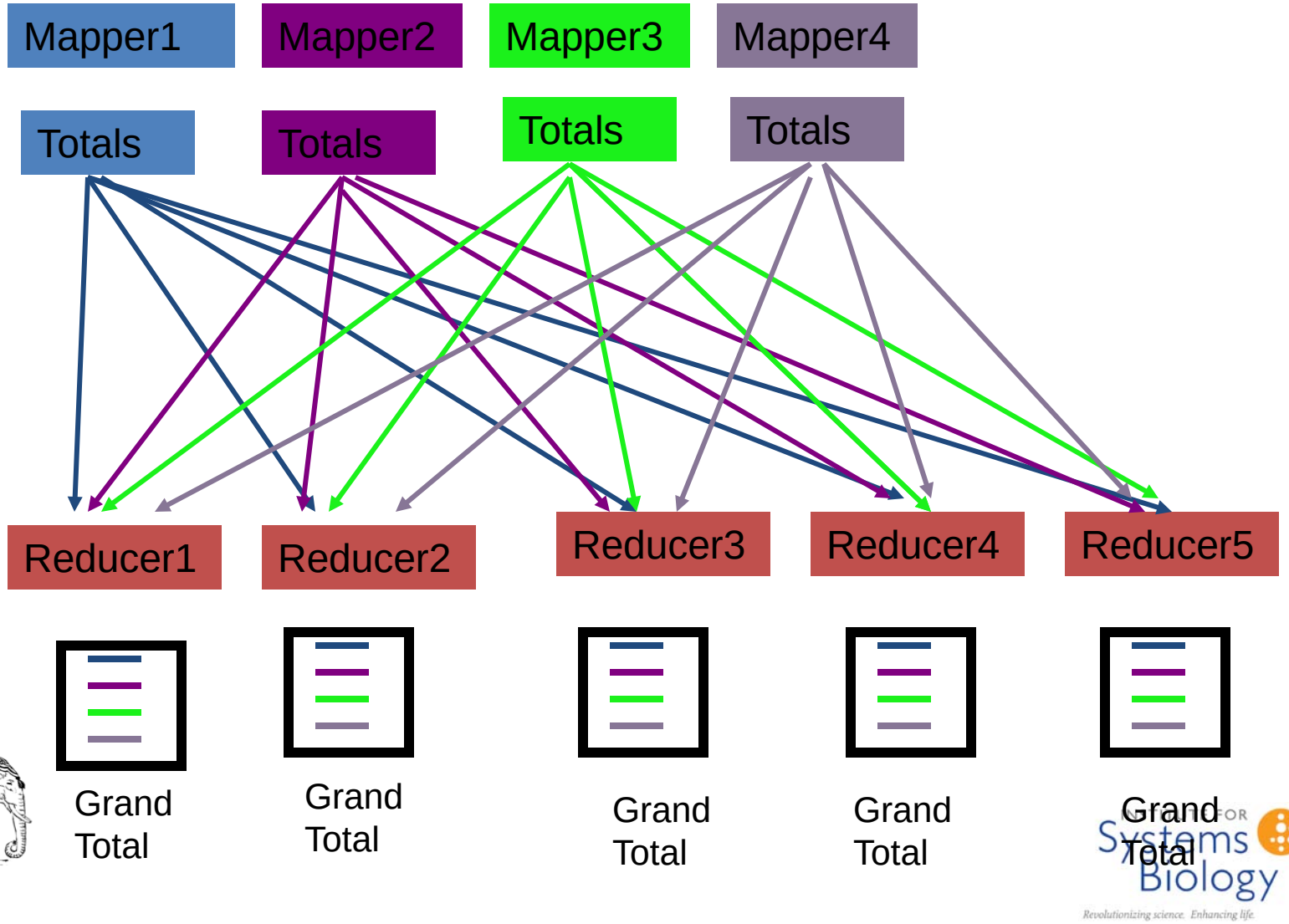
Getting Statistics to Reducers

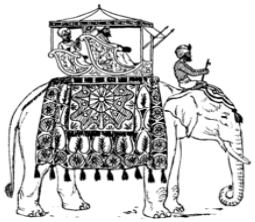
- For statistics to be used in reduction two conditions need to be met:
  1. Every reducer must receive statistics from every mapper
  2. All statistics must be received and processed before data requiring the statistics is handled

INSTITUTE FOR Systems Biology™
Revolutionizing science. Enhancing life.

Broadcast

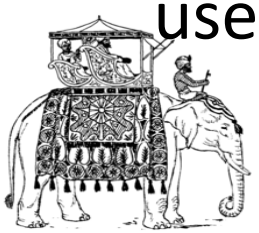
Every Mapper sends its total to each reducer – reducer makes grand total – before other keys sent





Consolidators

- Many – perhaps most map reduce jobs take a very large data set and generate a much smaller set of output.
- While the large set is being reduced it makes sense to have each reducer write to a separate file part-r-00000, part-r-00001 … independently and in parallel.
- Once the smaller output set is generated it makes sense for a single reducer to gather all input and write a single output file of a format of use to the user.
  - These tasks are called consolidators.



INSTITUTE FOR Systems Biology
Revolutionizing science. Enhancing life.

Consolidators

- Consolidation is the last step
- Consolidators can generate any output a in any location and frequently write off the HDFS cluster
- A single Howdah job can generate multiple consolidated files – all output to a given file is handled by a single reducer

INSTITUTE FOR Systems Biology
Revolutionizing science. Enhancing life.

Consolidation Process

- Consolidation mapper assigns data to a output file Key is original key prepended with file path
- Consolidation Reducer receives all data for an output file and writes that file using a path.
- The format of the output is controlled by the consolidator.
- Write is terminated by cleanup or receiving data for a different file

INSTITUTE FOR Systems Biology
Revolutionizing science. Enhancing life.

Biological Problems –

Shotgun DNA Sequencing

- DNA is cut into short segments
- The ends of each segment is sequenced
- Each end of a read is fit to the reference sequence

reference

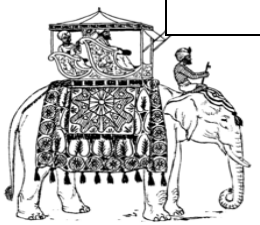
ACGTATTACGTACTACTACATAGATGTACAGTACTACAATAGATTCAAACATGATACA

Sequences with ends fit to reference

ATTACGTACTAC..... …….... ACAGTACTACAA

CGTATTACGTAC……..……ACTACAATAGATT

ACTACTACATA……..…….CAATAGATTCAAA



INSTITUTE FOR Systems Biology

Revolutionizing science. Enhancing life.

Processing

- Group all reads mapping to a specific region of the genome
- Find all reads overlapping a specific location on the genome

Location is Chromosome : location i.e.

CHRXIV:3426754

INSTITUTE FOR Systems Biology
Revolutionizing science. Enhancing life.

Single Mutation Detection

- Most sequences agree in every position with the reference sequence
- When many sequences disagree with the reference in one position but agree with each other a single mutation is suspected

SNP reference

ACGTATTACGTACTACTACATAGATGTAC
ACGTATTACGTAC
TTTTACGTACTACTA
GTTTTACGTACTAC
TTTTACGTACTACATAG
CGTATTACGTACTACTA

INSTITUTE FOR Systems Biology
Revolutionizing science. Enhancing life.

Deletion Detection

- Deletes are detected when the ends of a read are fitted to regions of the reference much further apart than normal
- The fit is the true length plus the deleted region

reference

ACGTATTACGTACTACTACATAGATGTACAGTACTACAATAGATTCAAACATGATACAACACACAGTA

actual deletion

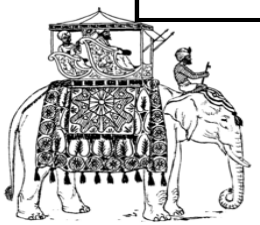
ACGTATTACGTACTAC|TCAAACATGATACAACACACAGTAAGATAGTTACACGTTTATATATATACC

fit to actual

ATTACGTACTAC..... ? ? ? ..TACAACACACAG

reported fit to reference

ATTACGTACTAC.....TACAACACACAG



INSTITUTE FOR Systems Biology

Revolutionizing science. Enhancing life.

Performance

- Platforms
  - Local – 4 Core Core2 Quad 4Gb
  - Cluster – 10 node cluster running Hadoop - 8 core per node 24Gb Ram 4 TB Disk
  - AWS –small node cluster (nodes specified) – 1gb virtual

INSTITUTE FOR Systems Biology
Revolutionizing science. Enhancing life.

Data

Platform		Task	Timing
Local – 1 cpu		200 M	15 min
Local – 1 cpu		2 GB	64 min
10 node cluster		2 GB	1.5 min
10 node cluster		1 TB	32 min
AWS 3 small		2 GB	7 Min
AWS 40 small		100GB	800 Min

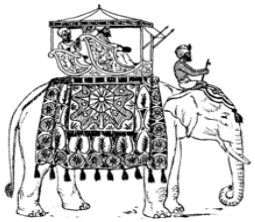
INSTITUTE FOR Systems Biology™
Revolutionizing science. Enhancing life.

Conclusion

- Howdah is useful for tasks where:
  - a large amount of data is processed into a much smaller output set
  - Multiple analyses and outputs are desired for the same data set
  - The format of the output file is defined and cannot simply be a concatenation
  - Complex processing of input data is required sometimes including broadcast global information

INSTITUTE FOR Systems Biology
Revolutionizing science. Enhancing life.

Questions



INSTITUTE FOR Systems Biology™
Revolutionizing science. Enhancing life.

Critical Elements

- Keys are enhanced by prepending a task specific ID
- Broadcast is handled by prepending an id that sorts before non-broadcast ids
- Consolidation is handled by prepending a file path and using a partitioner which assures that all data in one file is sent to the same reducer.

INSTITUTE FOR Systems Biology™
Revolutionizing science. Enhancing life.