Summary

University of Reading

- Goal (user-perspective): Optimise the time-to-solution
- Runtime of queries/scripts is the main contributor
- Computation in big data clusters is usually over-dimensioned
- Understanding a few HW throughputs help to assess the performance
- Linear scalability of the architecture is the crucial performance factor
- Basic performance analysis
  1. Estimate the workload
  2. Compute the workload throughput per node
  3. Compare with hardware capabilities
- Error model predicts runtime if jobs must be restarted
- GreySort with Spark utilises I/O, communication is good
- Computation even with Spark is much slower than Python
- Different big data solutions exhibit different performance behaviours