HADOOP IN PRACTICE 2014 pdf

1: Book Review: Hadoop in Practice â€“ Randal Scott King

Hadoop in Practice, Second Edition provides over tested, instantly useful techniques that will help you conquer big data, using Hadoop. This revised new edition covers changes and new features in the Hadoop core architecture, including MapReduce 2.

Understanding object models and storage formats 3. Parquet and the Hadoop ecosystem 3. The importance of output committing 3. Organizing and optimizing data in HDFS 4. Directory and file layout 4. Compacting Technique 28 Using filecrush to compact data Technique 29 Using Avro to store multiple small binary files 4. Atomic data movement 4. Moving data into and out of Hadoop 5. Key elements of data movement 5. Moving data into Hadoop 5. Moving data out of Hadoop 5. Applying MapReduce patterns to big data 6. Joining Technique 54 Picking the best join strategy for your data Technique 55 Filters, projections, and pushdowns 6. Map-side joins Technique 56 Joining data where one dataset can fit into memory Technique 57 Performing a semi-join on large datasets Technique 58 Joining on presorted and prepartitioned data 6. Reduce-side joins Technique 59 A basic repartition join Technique 60 Optimizing the repartition join Technique 61 Using Bloom filters to cut down on shuffled data 6. Data skew in reduce-side joins Technique 62 Joining large datasets with high join-key cardinality Technique 63 Handling skews generated by the hash partitioner 6. Sorting Technique 64 Implementing a secondary sort 6. Total order sorting Technique 65 Sorting keys across multiple reducers 6. Sampling Technique 66 Writing a reservoir-sampling InputFormat 6. Utilizing data structures and algorithms at scale 7. Modeling data and solving problems with graphs 7. Modeling graphs Technique 67 Find the shortest distance between two users 7. Tuning, debugging, and testing 8. Measure, measure, measure 8. Common inefficiencies in MapReduce jobs Technique 72 Viewing job statistics 8. Shuffle optimizations Technique 76 Using the combiner Technique 77 Blazingly fast sorting with binary comparators Technique 78 Tuning the shuffle internals 8. Reducer optimizations Technique 79 Too few or too many reducers 8. General tuning tips Technique 80 Using stack dumps to discover unoptimized user code Technique 81 Profiling your map and reduce tasks 8. Accessing container log output Technique 82 Examining task logs 8. Accessing container start scripts Technique 83 Figuring out the container startup command 8. MapReduce coding guidelines for effective debugging Technique 85 Augmenting MapReduce code for better debugging 8. Testing MapReduce jobs 8. Essential ingredients for effective unit testing 8. Integration and QA testing 8.

2: Manning | Hadoop in Practice

Hadoop in Practice collects 85 battle-tested examples and presents them in a problem/solution format. It balances conceptual foundations with practical recipes for key problem areas like data ingress and egress, serialization, and LZO compression.

Reading List In his new book Hadoop in Practice. Second Edition , Alex Holmes provides a comprehensive guide for Hadoop developers on leveraging Hadoop capabilities. Unlike the majority of Hadoop books describing basic Hadoop features, this one assumes that you are already familiar with them and discusses how to make the best of them in practice. Coupled with over a hundred practical recipes of writing Hadoop implementations, the content of the book is indispensable resource for Hadoop professionals. The book is organized in 10 chapters divided into 4 parts: It also provides a gentle introduction to YARN â€" a new resource manager introduced in Hadoop 2, which allows one to simultaneously run multiple different software stacks on top of a Hadoop cluster. It also covers setting up a single-node Hadoop cluster for experimenting with the code provided in the book. Related Vendor Content Related Sponsor Enhance your end-user experience by optimizing your application performance. Get a holistic view of your application behavior with Site24x7. It starts with chapter 3, describing different data serialization formats, which is one of the fundamental properties of data storage. The chapter provides pros and cons of different serialization methods and code samples for using them in MapReduce, Hive and Pig applications. Chapter 4 covers data organization in HDFS including data partitioning approaches, directory structures, etc. It also covers using compression for optimizing data storage and the impact compression has on data splitability. Additionally, this chapter covers using Scoop for data transfer to and from relational databases and integration with HBase. Finally, automation of data transfer using cron jobs and Oozie. Chapter 6 covers implementing common big data process patterns including joining, sorting and sampling. It describes approaches to implementation of these patterns and provides code samples for their implementation. Chapter 7 looks at more advanced data structures and algorithms that can be used for big data processing. It starts with using graph processing for solving common problems like shortest distance, friends of friends and PageRank, sketching their implementation in MapReduce. It then demonstrates using Bloom filters for effective membership queries and HyperLogLog for count estimations, showing how these structures can be calculated and leveraged by MapReduce implementations. Finally, Chapter 8 describes approaches and best practices for debugging, testing and tuning MapReduce applications. InfoQ had a chance to interview Holmes. Do you consider these as auxiliary? XML and JSON are constrained when it comes to areas such as splitability and schema evolution, which is why Avro and Parquet are compelling alternatives. While columnar data formats in general and Parquet specifically are wildly used for real-time SQL engines, like Hive, Impala, Drill, etc. I do not really see them being widely used in MapReduce. Why do you consider them useful here? For the same benefits that columnar formats are useful in SQL engines â€" the ability to use projection and predicate pushdowns to optimize reads in your jobs. You can see them in action in technique 24 in chapter 3. Any reason for this? What do you consider to be a typical MapReduce application? As a result MapReduce continues to excel at ETL-like workloads that require bulk, batch methods to move and transform data. Is it MapReduce, Giraph, something else? Although Chapter 7 provides a good description on computing both Bloom filters and HyperLogLog, it does not provide any best practices on using them. Can you elaborate on these structures? They are both probabilistic data structure that optimize for space at the expense of accuracy. Bloom filters are incredibly useful for filtering operations, an example of which is provided in technique 61 in chapter 6. HyperLogLog are essential ingredients when building Lambda architectures and you need the ability to provide distinct element counts over extremely large datasets. There have been a few publications questioning the usefulness of MRUnit for testing MapReduce applications. What is your opinion on that? Currently SQL is often considered the most important Hadoop technology, effectively turning Hadoop into a humongous database. Do you share this opinion? It opens-up the audience of Hadoop to data scientists and analysts, providing them the tools needed to quickly craft sophisticated queries to dissect and pick out their data. Do you consider Spark an improved version of MapReduce? I think Spark has a very promising future, and is already making key inroads into areas where MapReduce used to be the only solution. One of the hardest decisions in picking a tool chain is ensuring that it works reliably and predictably over the huge datasets we have in production, and does so in a way that plays nice with other users and products running in our systems. I think Spark is still early in its lifecycle, and I look forward to improved administration and profiling capabilities to help with tuning our applications. How do you see the future of the Hadoop ecosystem? I believe security and multi-tenancy are two areas where we need continued focus. Spark and Tez are also exciting as they move beyond MapReduce and help us engineer more efficient data pipelines. About the Book Author Alex Holmes is a senior software engineer with over 15 years of experience developing large scale distributed Java systems. For the last five years he has gained expertise in Hadoop solving Big Data problems across a number of projects. He has presented at the JavaOne and Jazoon conferences.

3: Hadoop in Practice | PDF Free Download

Kalyan Hadoop Training in Hyderabad @ ORIEN IT, Ameerpet, , Hadoop in Practice, hadoop training in hyderabad, spark training in hyderabad.

4: Hadoop in Practice, 2nd Edition : Books

Hadoop in Practice, Second Edition provides over tested, instantly-useful techniques that will help conquer big data, using Hadoop. This revised new edition covers changes and new features in the Hadoop core architecture, including MapReduce 2.

5: Hadoop in Practice by Alex Holmes

You sit on a big pile of data and want to know how to leverage it in your company? Interested in use-cases, examples and practical demos about the full Hadoop .

6: Manning | Hadoop in Practice, Second Edition

Annotation 'Summary Hadoop in Practice' provides over tested, instantly useful techniques that will help you conquer big data, using Hadoop.

7: Hadoop in Practice by Alex Holmes (, Paperback) | eBay

Â© Manning Publications Co. All rights reserved. Powered by JForum , Â© JForum TeamJForum , Â© JForum Team.

8: Hadoop in Practice - GeekBooks - Free Tech PDF eBook Library

application, the goal of this article is to demystify how MapReduce works in Hadoop 2. Dissecting a YARN MapReduce application Architectural changes had to be made to MapReduce to port it to YARN.

9: Hadoop in Practice, Second Edition [Book]

Note: Citations are based on reference standards. However, formatting rules can vary widely between applications and fields of interest or study. The specific requirements or preferences of your reviewing publisher, classroom teacher, institution or organization should be applied.

HADOOP IN PRACTICE 2014 pdf

Guide to federal pharmacy law reiss The Ultimates, Vol. 1 Implementation of Functional Languages: 9th International Workshop, Ifl97 : St. Andrews, Scotland, Uk Sep Teachings of yoga But Ill be back again V. 3-4. The history of Pendennis. Perumalu a/l Kandan and Ibrahim Bin Ali 62 How to discover your personal painting style Babbits and Bohemians Nonverbal vocal communication Tiny for windows The Pressures Off College physics young and geller 9th edition V. 2. The case of Africa Aristotelian/scholastic hylomorphism and the rise of mechanism Ancient monuments of Orkney Stories of Notable Women for Readers Theatre (Teacher Ideas Press) Bacons last captain Fashion Book, The Mini Edition Freedom Flyers Big Color Book Environment and embodiment in early modern England Whitehead Encyclopedia Of Deer Handbook of bird biology Mrs. Hill B.H. Fairchild Eric jerome dickey an accidental affair Communication Technologies Interference powder Victoria danann journey man Steviol glycosides Mak.Frankfurt (Prestel Museum Guides) Melina Nicolaides Broadcast News Writing, Reporting, and Producing The Sunbridge over the River V. 3, pt. 2. Kates, M. Techniques of lipidology. Counseling and psychotherapy theories and interventions 6th edition Inheriting the earth Yuck, a love story Theory of nuclear structure Interventions and policy shifts Theoretical aspects of mainly low dimensional magnetic systems