# Sessions Abstract

## Keynotes Addresses

### Hadoop and the Future of Cloud Computing

***Dr. Todd Papaioannou - Vice President, Cloud Architecture, Yahoo!***

#### Abstract

In this keynote, Todd will discuss why Yahoo! invests in Hadoop and Open Source, how Yahoo! uses Hadoop to power some of its most innovative technology products, the State of the Hadoop community and the future of BigData and Cloud Computing.

#### Speaker Bio

Dr. Todd Papaioannou is vice president, cloud architecture for Yahoo!’s global cloud computing group. The company uses its cloud computing infrastructure to power nearly all of its consumer and advertiser experiences. Yahoo!’s cloud stores, analyzes and processes hundred of petabytes of data and content, delivers it in a reliable and scalable way, making it accessible to users around the world on a variety of platforms and form factors.

Prior to joining Yahoo!, Dr. Papaioannou served as vice president for architecture and emerging technologies at Teradata. He focused on product and architectural strategy across the entire Teradata product portfolio. His recent focus included defining and driving Teradata's initiatives in the cloud computing and virtualization spaces and launching the Teradata Developer Exchange. Previously he was the CTO of Teradata's client software group and served as the chief architect for the Teradata Viewpoint program from inception through the first two product releases. Prior to joining Teradata, Dr. Papaioannou was chief architect at Greenplum/Metapa. Dr. Papaioannou holds a PhD in artificial intelligence and distributed systems.

### Programming Abstractions for Smart Apps on Clouds

***Dr. D. Janakiram - Professor, Department of CSE, Indian Institute of Technology, Madras***

#### Abstract

Talk will focus on how the programming abstractions are moving from map-reduce (predominantly for file and keyword indexing) to Dryad (more of database applications) to programming Artificial Intelligent applications on Clouds. I could discuss programming these abstractions both on Hadoop and Edge Node File Systems (ENFS - research from our lab on Voluntary clouds).

#### Speaker Bio

Dr. D. Janakiram is currently Professor at the Department of Computer Science and Engineering, Indian Institute of Technology (IIT), Madras, where he heads and coordinates the research activities of the Software Systems Lab and Distributed Systems Lab. He is Founder of the Forum for Promotion of Object Technology, which conducts the National Conference on Object Oriented Technology (NCOOT) and Software Design and Architecture (SoDA) workshop annually. He is also principal investigator for a number of projects including the grid computing project at the Department of Science and Technology, Linux redesign project at the Department of Information Technology, and Middleware Design for Wireless Sensor Networks at the Honeywell Research Labs, Indo-German Project on Mobile Telemedicine Grid, Indo-Italian Project on Peer-Peer Semantic Search and Xerox Project on device grids and cloud. He is Program Chair for the 8th International Conference on Management of Data (COMAD).

### Exploring the Future IT Infrastructure, Cloud Included

***Sundara Nagarajan, Director of R&D, Hewlett-Packard India Software Operation***

#### Abstract

Cloud computing has caught the imagination of consumers, entrepreneurs, companies and governments. What makes cloud and associated technologies most attractive is the opportunity for radical business innovation and new business models. As the expectations are on the rise from users, it is important for engineers and researchers to understand the underlying issues in building and operating these mission-critical systems. This talk will be focused on exploring how the IT infrastructure technologies are changing in order to address the needs of information technology as it evolves. The speaker plans to touch upon topics in computer systems architecture, operating systems and data services for the IT infrastructure of the next generation. What are the new computing elements? How will they be interconnected and managed? How the data architecture and access mechanisms are changing? What are the challenging technical problems in realizing the infrastructure of the future? The author also plans to introduce OpenCirrus™, an open cloud-computing research test-bed designed to support research into the design, provisioning, and management of services at a global, multi-datacenter scale.

#### Speaker Bio

Sundara Nagarajan ("SN") is currently Director of R&D in the Storage Works Division, based in Bangalore, India. He leads an R&D team engineering products, in the broad technical domain of virtualized and scale-out storage.

Previously he was Distinguished Technologist and Director of R&D, responsible for R&D management of systems software components for enterprise-class servers and server management products. He has over 27 years of experience in Computer Systems R&D. He also serves as a Visiting Professor at International Institute of Information Technology, Bangalore.

SN holds graduate degree in Electrical Engineering from the University of Calicut and M.S. (By Research) from the Indian Institute of Technology, Madras. He is a Certified ScrumMaster, Senior Member of the IEEE and Member of the ACM.

### Federated HDFS

***Sanjay Radia – Cloud Architect, Yahoo!***

#### Abstract

Scalability of the NameNode has been a key struggle. Because the NameNode keeps all the namespace and block locations in memory, the size of the NameNode heap limits the number of files and also the number of blocks addressable. This also limits the total cluster storage that can be supported by the NameNode.

Federated HDFS allows multiple independent namespaces (and NameNodes) to share the physical storage within a cluster. This is enabled by the introduction of the notion of Block pools which is analogous to LUNs in a SAN storage system

This approach offers a number of advantages besides scalability: it can isolate namespaces of different applications improving the overall availability of the cluster. The Block pool **Abstract**ion allows other services (such as HBase) to use the block storage with perhaps a different namespace structure.

Applications prefer to continue to use a single namespace. Namespaces can be mounted to create such a unified view. A client-slide mount table provides an efficient way to do that, compared to a server-side mount table: it avoids an RPC to the central mount table and is also tolerant of its failure. The simplest approach is to have shared cluster-wide namespace; this can be achieved by giving the same client-side mount table to each client of the cluster. Client-side mount tables also allow applications to create a private namespace view. This is analogous to the per-process namespaces that are used to deal with remote execution in distributed systems.

#### Speaker Bio

Sanjay is the architect of the Hadoop project at Yahoo where it is in daily use for large clusters of several thousand machines. Previously he has held senior engineering positions at Cassatt, Sun Microsystems and INRIA where he developed systems software for distributed systems and grid/utility computing infrastructures. He has published numerous papers and holds several patents. Sanjay has PhD in Computer Science from University of Waterloo, Canada.

### Scaling Hadoop Applications

***Dr. Milind Bhandarkar - LinkedIn***

#### Abstract

Apache Hadoop makes it extremely easy to develop parallel programs based on MapReduce programming paradigm by taking care of work decomposition, distribution, assignment, communication, monitoring, and handling intermittent failures. However, developing Hadoop applications that linearly scale to hundreds, or even thousands of nodes requires extensive understanding of Hadoop architecture and internals, in addition to hundreds of tunable configuration parameters. In this talk, I illustrate common techniques for building scalable Hadoop applications, and pitfalls to avoid. I will explain the seven major causes of sub linear scalability of parallel programs in the context of Hadoop, with real-world examples based on my experiences with hundreds of production applications at Yahoo! and elsewhere. I will conclude with a scalability checklist for Hadoop applications, and a methodical approach to identify and eliminate scalability bottlenecks.

#### Speaker Bio

Dr. Milind Bhandarkar was the founding member of the team at Yahoo! that took Apache Hadoop from 20-node prototype to datacenter-scale production system, and has been contributing and working with Hadoop since version 0.1.0. He started the Yahoo! Grid solutions team focused on training, consulting, and supporting hundreds of new migrants to Hadoop. Parallel programming languages and paradigms has been his area of focus for over 20 years. He worked at the Center for Development of Advanced Computing (C-DAC), National Center for Supercomputing Applications (NCSA), Center for Simulation of Advanced Rockets, Siebel Systems, Pathscale Inc. (acquired by QLogic), and Yahoo!. Currently, he works on distributed data systems at LinkedIn Corp.

## Platform Track

### Hadoop - NextGen

***Sharad Agrawal – Technical, Yahoo!***

#### Abstract

Hadoop has come a long way from running on 20 nodes prototype to 4000 nodes production cluster in last couple of years. However there are certain challenges in maintaining and taking the current Hadoop Map-Reduce platform to the next level. In particular, the Map-Reduce JobTracker needs an overhaul to address certain technical deficiencies in its memory consumption and concurrency model to make it more scalable and performant. In this talk I will present a new architecture for Hadoop Map-Reduce which tackles these challenges. Also as part of this architecture, I will talk about generic resource scheduler which enables to run programming paradigms other than Map-Reduce on the shared Hadoop cluster.

#### Speaker Bio

Sharad is a Hadoop Committer and a member of Hadoop Project Management Committee. He is working with Yahoo since 2007. At Yahoo, he has worked in search advertising domain and been part of various platform development including Hadoop Map-Reduce and Apache Avro project. Prior to joining Yahoo, he has worked in the area of vertical search at AOL. Sharad has a BTECH from IIT Delhi.

### Pig, Making Hadoop Easy

***Alan Gates - Pig Architect, Yahoo!***

#### Abstract

Pig is a platform for analyzing large data sets. It consists of a high-level language, Pig Latin, for expressing data analysis programs, coupled with infrastructure for evaluating these programs atop Hadoop’s MapReduce platform. This talk will review the basic features of Pig, discuss recent interesting additions to the system as well as current work being done, talk about Pig performance, and consider areas for future development and research.

#### Speaker Bio

Alan Gates is an architect in the grid team at Yahoo. He has been a committer on the Pig project since 2007. He has been developing database and data processing technology for the last twelve years, including eight years at Yahoo dealing with storage and query engines for petabyte sized data sets.

### Data on Grid (GDM)

***Venkatesh S – Sr. Technical Yahoo!***

#### Abstract

#### Speaker Bio

### Hive Evolution

***Namit Jain - Facebook***

#### Abstract

Hive is an open source, peta-byte scale date warehousing framework built on top of Hadoop. Hive supports queries expressed in a SQL-like declarative language - HiveQL, which are compiled into MapReduce jobs that are executed using Hadoop. In addition, HiveQL enables users to plug in custom map-reduce scripts into queries.

The language includes a type system with support for tables containing primitive types, collections like arrays and maps, and nested compositions of the same. The underlying IO libraries can be extended to query data in custom formats. Hive also includes a system catalog - Metastore that contains schemas and statistics, which are useful in data exploration, query optimization and query compilation.

In this presentation we will be talking in more detail about Hive, the motivations behind it, its evolution into an apache top-level project, the work in progress and the challenges ahead. We will also be briefly touching upon its usage in Facebook, where the Hive warehouse contains tens of thousands of tables and stores over 30PB of data and is being used extensively for both reporting and ad-hoc analyses by more than 400 users per month.

#### Speaker Bio

Namit Jain is the chair of Hive at apache. He has been with Facebook’s data-infrastructure group for about 2.5 years. Before that, he worked for over 10 years at Oracle on streaming technologies, XML, replication and queuing.

### Making Hadoop Secure

***Devaraj Das - Sr. Technical, Yahoo***

#### Abstract

Hadoop, until recently, would trust any user based on who he says he is. This is clearly not enough in large companies where they have Hadoop instances storing sensitive data (like financial, revenue, etc.), and where these instances are being used by many users and from potentially different groups. In this talk, I will cover the security threats in Hadoop in the various communication paths (in Hadoop Distributed File System, MapReduce, and the client components). I will present the solutions we designed for each of them. I’ll also cover briefly the security solutions to do with external services like Oozie & HDFSProxy talking with Hadoop.

#### Speaker Bio

Devaraj Das is an Apache Hadoop Committer and a member of the Apache Hadoop PMC. He is a senior engineer in the Cloud Platform Group at Yahoo! Sunnyvale, California.

### GridSim: A Benchmark Suite for Hadoop and Pig

***Ranjit Mathew - Technical, Yahoo***

#### Abstract

Hadoop and Pig are widely used within Yahoo for data preparation and analysis. While Hadoop allows scalable and reliable computing using commodity hardware, Pig allows developers to prepare and process data using Hadoop without having to write tedious MapReduce programs themselves. Measuring and analyzing their performance is important for improving utilization of existing hardware and for capacity-planning. GridSim is a suite of benchmarks developed for this purpose, comprising GridMix3 and PigMix2. It has been used successfully for certifying Hadoop and Pig releases, reproducing issues seen in production clusters, ensuring that patches do not cause performance regressions, identifying performance bottlenecks, etc.

#### Speaker Bio

Ranjit works in the Hadoop Engineering team of Yahoo! R&D, Bangalore. He has over 14 years of experience working in the computer industry, including working for companies like Oracle and IBM. He graduated in Computer Science and Engineering from the Indian Institute of Technology (IIT), Kanpur.

### Oozie - Workflow for Hadoop

***Andreas Neumann - Architect, Yahoo***

#### Abstract

Oozie is a workflow engine for defining, scheduling, executing, and monitoring complex processes of dependent Grid tasks, such as map/reduce jobs, Pig scripts or HDFS operations; and it is extensible to other types of tasks, such as Hive. Oozie has found wide adoption at Yahoo!, and today it controls hundreds of production processes on Yahoo!'s grids. Oozie is open-sourced in Github and approaching its 3rd major release. This talk will introduce the main features of Oozie - workflows and coordinators - and give an overview of upcoming new features, such as coordinator bundles and authentication.

#### Speaker Bio

Andreas has been at Yahoo! since 2008, where he has worked on Web crawling and content platforms before joining the Grid team. Andreas was involved in the design of Oozie from the early days. Before joining Yahoo, Andreas worked on Enterprise Search at IBM. He received his PhD in Computer Science from the University of Trier, Germany, for his work on querying structured documents.

## Application Track

### Online content optimization using Hadoop

***Dr. Shail Aditya Gupta - Yahoo!***

#### Abstract

Content optimization is about delivering the *right* content to the *right* user at the *right* time. On Yahoo's today module on front page, every 5 minutes we have 32,000 different variations of that module which we serve to 500 million users. I would describe this challenge of content optimization at scale and how Yahoo! is leveraging the power of Hadoop stack to conquer this challenge. Hadoop, along with Hive and Hbase, is the technology that enables mass content personalization for Yahoo! users. I will cover the high level architecture and the various modeling challenges of this content optimization engine.

#### Speaker Bio

Dr. Shail Aditya is a Senior Architect in the Cloud Platform Group in Yahoo! SDC, Bangalore. He is currently working with the Content Optimization team to rank and serve trending content in near real time. His background is from the field of high performance compilers and architectures and electronic design automation (EDA). Shail received his B.Tech. in Computer Science and Engineering from the Indian Institute of Technology, Delhi, and his S.M., E.E. and Ph.D. in Electrical Engineering and Computer Science from MIT.

### Making Hadoop Enterprise ready with Amazon Elastic Map/Reduce

***Simone Brunozzi– Technical Architect, Amazon***

#### Abstract

As the demand for cloud-based analysis of large data sets explodes, customers of Amazon Web Services have wanted to leverage Hadoop in the cloud. Amazon Elastic MapReduce manages the complex and time-consuming set-up, operation and tuning of both Hadoop clusters and the compute capacity upon which they sit. A user can instantly spin up large Hadoop job flows and begin processing within minutes. Over the last year AWS has worked with current users to develop new features that make it even easier to execute Hadoop applications in the cloud. In this session, we will review lessons learned from those current users, discuss recent improvements to the Amazon Elastic MapReduce, and look at key features coming in the near future. In addition, we will discuss developments in the ecosystem in which exciting offerings developed on top of Amazon Elastic MapReduce have made it an even more compelling solution for enterprise Big Data analytics.

#### Speaker Bio

Simone Brunozzi is a truly passionate technologist, involved in IT and computing since an early age.  
He joined Amazon.com in March 2008 in the role of Amazon Web Services Technology Evangelist for Europe, to raise awareness about Cloud Computing and help developers build applications.  
In January 2010 he relocated to Singapore, becoming the first AWS Technology Evangelist for the APAC region.

### Hadoop Avatar at eBay

***Srinivasan Rengarajan, Mohit Soni - Senior Staff Engineer, eBay Chennai***

#### Abstract

In the last 15 years, eBay grew from a simple website for online auctions to a full-scale e-commerce enterprise. It processes petabytes of data each day to solve a gamut of complex problems like product search, product recommendation, fraud detection, business intelligence etc. With the ever increasing complexity of problems and exponentially increasing data, eBay has turned to Hadoop for a scalable and reliable solution.

In this session we will talk about the Hadoop ecosystem at eBay. We will provide a high level overview of deployment and internal user base. We will also briefly talk about the Mobius platform which was internally developed at eBay. It is a wrapper around Hadoop which is used as a self service tool for click stream analysis. If time permits we will present the results of some real life problems being tackled at eBay using Hadoop.

#### Speaker Bio

Srinivasan Rengarajan works as Technical Architect at eBay Chennai. Prior to eBay, he worked for Informatica (analytics/metadata servers) and in Sun Bangalore as a Staff Engineer, helped pioneer the first of its kind Open Source Master Data Management Project mural. He Has Spoken at various conferences such as Sun Tech-Days and JavaOne.

Mohit Soni started with e-commerce giant eBay as an Intern in 2010 and worked with eBay Research Labs on distributed computing, search, IR etc. He is currently working with eBay as a Senior Software Engineer. He graduated in Computer Science from VIT University, Vellore.

### Feeds processing at Yahoo! , One Hadoop, one platform, 2 systems

***Jean-Christophe Counio - Yahoo!***

#### Abstract

One of the first applications of Hadoop used in production at Yahoo! has been Pacman, a system meant to process large sized feeds (millions of records). With time, this system has been more and more used to ingest small feeds, but didn’t really answered well the new requirements. This lead to the creation of a new system, Pepper, able to scale a large number of small feeds with very low overhead, using Hadoop in a non-traditional way.

We’ll show in this presentation the different steps we went through to build the feeds processing platform, the design of both Pacman and Pepper, and how we now are able to support efficiently the whole spectrum of feeds Yahoo! receives, while keeping a unique Hadoop cluster to scale the load. We’ll detail the contributions brought to Hadoop in the last years, then provide production numbers and examples of processing.

#### Speaker Bio

Jean-Christophe Counio has lead different projects at Yahoo! and joined the feeds platform team in 2009 as an architect.

### Hadoop 101

***Basant Verma – Sr. Technical, Yahoo!***

#### Abstract

This talk will focus on basic concepts of developing Hadoop applications and walk through various optimizations techniques and best practices for improving performance of Hadoop applications.

#### Speaker Bio

Basant Verma is Grid Solutions Architect at Yahoo where he helps Yahoo internal groups in designing, building and deploying large Data crunching applications using Hadoop.

### Searching Information inside Hadoop Platform

***Abinasha Karana - Director of Technology and Co-Founder, Bizosys Technologies***

#### Abstract

In the context of storing vast number of documents on HDFS and large volumes of Records on HBase, how does one find information from across this huge combination of structured data and unstructured information? In this presentation you will hear about how Bizosys began with typical search solutions such as Lucene. Preliminary investigations revealed scalability issues similar to Relational Databases for records. Subsequently, Bizosys developed a distributed, real-time search engine whose Index is stored and served out of HBase. The presentation will cover some of the key architectural learnings from Hadoop technology platform pertaining to issues such as - ways to move processing closer to data; Reducing IPC calls; Balancing network vs. CPU vs. I/O and memory, etc. to find/read data faster from HBase and Hadoop.

#### Speaker Bio

Abinash is currently the Director - Technology and Co-Founder at Bizosys Technologies Pvt Ltd, Bangalore. Bizosys Technologies is a Bangalore based software product company that caters to scale Big Data, and Business-IT collaboration. Prior to Bizosys, Abinasha was chief architect and co-founder of Drapsa Technologies Pvt Ltd in 2007, which was merged with another company with an interest in retail solutions. In his career with Infosys Technologies, Bangalore, he was involved in various initiatives such as - starting the Infosys Mobility Solutions and Enterprise Search practice. One of his architected projects featured in the 2004 InfoWorld 100. At Infosys, Abinasha's efforts were recognized with awards such as Infoscion of the Week, North America All Star. Abinasha graduated in Engineering from NIT, Rourkela and is PMP, TOGAF certified.

### Data Integration on Hadoop

***Sanjay Kaluskar – Senior Architect, Informatica Corporation***

#### Abstract

Data fragmentation is a harsh reality that many enterprises have to deal with as they strive to get a global picture. Data gets fragmented as an enterprise grows, due to the existence of many users and applications (ERP, CRM, email, home-grown applications, etc.); these are well-recognized challenges in creating data warehouses and data marts. To integrate data from many diverse sources is a non-trivial problem: one must understand the API of the data source, the data schema, vendor and domain specific formats and often transform and correlate the data from multiple sources. Another challenge is data quality; often data is incomplete, inconsistent or inaccurate. These same challenges need to be tackled when users try to use Hadoop for analytics, trend-prediction, better services, etc.

Informatica Corporation is the world’s number one independent provider of data integration and data quality software. Organizations around the world gain a competitive advantage in today’s global information economy with timely, relevant and trustworthy data for their top business imperatives. More than 4,200 enterprises worldwide rely on Informatica to access, integrate and trust their information assets held in the traditional enterprise, off premise and in the Cloud.

I will describe Infadoop, an integration of Informatica and PIG currently implemented as a prototype. The prototype allows PIG users to leverage Informatica’s data integration and data quality features such as connectivity and a rich set of transformations. PIG users can also re-use existing data integration or data quality logic (mapplets) created using Informatica’s designer. The prototype also allows the conversion of existing Informatica logic (mappings) into PIG which can easily allow an Informatica user to leverage PIG (and Hadoop).

I will describe the overall functionality, implementation approach and challenges, which might be relevant to other PIG users. The PIG and Hadoop development community can be aware of how Informatica is contributing to the ecosystem, and expanding the user base. Of course, any feedback from the community will help tremendously as we move towards productization.

#### Speaker Bio

Sanjay Kaluskar works as a senior architect at Informatica. He did a B.Tech. in computer science from IIT Kanpur and an M.S. in computer science from Univ. of Texas at Austin. Prior to Informatica he has worked at Yahoo, Oracle and IIT Mumbai. Sanjay has 17 years of industry experience in the areas of systems software, database systems, application servers, Internet applications and data integration tools.

## Research Track

### Middleware Frameworks for Adaptive Executions and Visualizations of Climate and Weather Applications on grids

***Dr. Sathish Vadhiyar - IISc Bangalore***

#### Abstract

Online remote visualization and steering of critical climate and weather applications like cyclone tracking are essential for effective and timely analysis by geographically distributed climate science community. We have developed an integrated user-driven and automated steering framework for simulations, online remote visualization, and analysis for critical weather applications. Our framework provides the user control over various application parameters including region of interest, resolution of simulations, and frequency of data for visualization. We have also developed middleware framework for efficient execution of long-running climate modeling applications on grids. This talk will focus on these middleware efforts.

#### Speaker Bio

Dr. Vadhiyar is assistant Professor in SERC, IISc since November 2003. Prior to that, from 1999-2003, I was in Innovative Computing Lab of Dr. Jack Dongarra, Computer Science Department, University of Tennessee, USA, first as a PhD student and then as a senior research associate.

### Comparison between Extension of Fairshare Scheduler and a Novel SLA based Learning Scheduler in Hadoop

***Dr. G Sudha Sadasivam , N Priya - PSG Tech, Coimbatore***

#### Abstract

This research project mainly aims at a comparison between extensions of schedulers in hadoop. It is mainly categorized and designed as two enhancements to scheduling and a comparison is to be done between the two. The first part is an extension to fairshare scheduler in hadoop and second part is the design and development a novel learning scheduler which is based on SLA (Service Level Agreement).

Extension to Fairshare Scheduler in Hadoop: - It is based on Large Job First with Small Job Backfilling (LSFB), where, the large jobs are brought to the beginning of the queue. It also incorporates the idea of backfilling of small jobs in between two large jobs in delay time.

A SLA Based novel learning scheduler for Hadoop: - This part mainly aims at design and development of a learning scheduler for Hadoop. It mainly aims at providing user level response and resource utilization. The JobTracker schedules the job based on Job traces history or from user given requirements.

#### Speaker Bio

### VirtPerf: A Capacity Planning Tool for Virtual Environments

***Dr. Umesh Bellur - Associate Professor, IIT Bombay***

#### Abstract

Several applications in the "physical" world are being consolidated in "virtual" environments using different virtualization technologies. An important criterion for this exercise is to understand potential resource utilization/requirements and performance levels these applications will require/achieve in virtual environments. Empirical evidence of this can be gotten by benchmarking the application's performance in a controlled manner on a virtual environment of the kind it will eventually be deployed on. These measurements can be used for a variety of purposes from virtual machine capacity planning to building sophisticated performance models that can be used to predict performance for loads that cannot be practically tested. In this paper, we present VirtPerf, an integrated workload generator and measurement tool to capture resource utilization levels and performance metrics of applications executing under controlled circumstances in virtualized environments. The tool aims to provide comprehensive measurement-based analysis for applications in different virtualization settings. Additionally, a configurable workload generator can be used to profile applications under different load conditions. We present the detailed design of VirtPerf and and a comprehensive empirical study to demonstrate its correctness and capabilities.

#### Speaker Bio

After his Ph.D., Umesh went to work in the industry where he has helped establish distributed object standards such as CORBA with OMG and J2EE with JCP . He worked for over 10 years at Oracle Corporation, Teknekron Communication Systems and Covad communications after which he helped found a startup in Silicon Valley called Collation Inc. in 2001 that was subsequently acquired by the IBM Tivoli group in 2005. He moved back to India and joined IIT Bombay as an Associate Professor in the Department of Computer Science and Engineering where he is currently. He is the recipient of the 2006 IBM Faculty Award in Autonomic Computing and the SAP Research and Innovation Award for QoS based overlay routing in 2008. His areas of research include: virtualization and cloud computing, adaptability in service oriented environments and autonomic computing techniques for distributed component based applications. This includes middleware design for different kinds of distributed systems including wireless sensor networks as well as QoS models for such environments.

### Scheduling in MapReduce using Machine Learning Techniques

***Dr. Vasudev Varma - IIIT Hyderabad***

#### Abstract

The MapReduce paradigm has become a popular way of expressing distributed data processing problems that need to deal with large amount of data. Despite the popularity and stability of Hadoop, it presents many opportunities for researching new resource management algorithms. Admission Control, Task Assignment and Scheduling, Data Local Execution, Speculative Execution and Replica placement are some of the key challenges involved in resource management in Hadoop. We address two of the above problems: Admission Control and Task Assignment. Admission controller, or the module that handles admission control, follows a learning based opportunistic algorithm that admits MapReduce jobs only if they are unlikely to cross the overload threshold set by the service provider. Our algorithm tries to maximize the utility from the service provider’s point of view and meets deadlines negotiated by users in more than 80% of cases. In another approach, our Task Assignment algorithm makes use of job properties for scheduling its tasks on the cluster. These job properties are tested using a Naive Bayes classifier to obtain a task compatible with the tasks currently running on a particular node.

#### Speaker Bio

Vasudeva Varma is a faculty member at International Institute of Information Technology, Hyderabad Since 2002. His research interests include search (information retrieval), information extraction, information access, knowledge management, cloud computing and software engineering. He is heading Search and Information Extraction Lab and Software Engineering Research Lab at IIIT Hyderabad. He is also the chair of Post Graduate Programs since 2009.

He published a book on Software Architecture (Pearson Education) and over seventy technical papers in journals and conferences. In 2004, he obtained young scientist award and grant from Department of Science and Technology, Government of India, for his proposal on personalized search engines. In 2007, he was given Research Faculty Award by AOL Labs.

He was visiting professor at UPV, Valencia, Spain (Summer 2007), UBO, Bretagne, France (Summer 2009) and Language Technologies Institute, CMU, Pittsburgh, USA (Summer 2010)

He obtained his Ph.D. from the Department of Computer and Information Sciences, University of Hyderabad in 1996. Prior to joining IIIT Hyderabad, he was the president of MediaCognition India Pvt. Ltd and Chief Architect at MediaCognition Inc. (Cupertino, CA). Earlier he was the director of Engineering and research at InfoDream Corporation, Santa Clara, CA. He also worked for Citicorp and Muze Inc. in New York as senior consultant.

### Adaptive parallel computing over distributed military computing infrastructures

***Dr. Rituraj Kumar - Director, DRDO Labs, Bangalore***

#### Abstract

The net centric paradigm of warfare would require increased dependence on algorithmic and mathematical solutions to battlefield situations. The scaling of the projected compute requirements would require an extent of parallelization to achieve expected response times. However, the tactical contexts would not be amenable to deployment of either large clusters in the tactical arena, nor would a back-haul to a cluster be feasible in most contexts. Therefore the solution would reside in leveraging the spare compute capacity of a large number of semi-autonomous compute devices that constitute the information grid. The presentation highlights some of the research directions in this space, leveraging various strategies of distributed computing and distributed file systems.

#### Speaker Bio

Dr. Rituraj Kumar received his M.Tech from IISc Bangalore, and his Ph.D. from IIT Delhi. He leads a group involved in R&D activities related to distributed, parallel and middleware computing. He has been involved in the design and development of Command and Control systems, especially focusing on the issues of Information architectures, design patterns and system integration. He is also involved in the area of Information security, where his primary interests are in Survivable Network Analysis and Trusted Computing Platforms. He has published about ten papers in archived journals and in national and international conferences

He has received the Laboratory Scientist of the year award 2004, National Science Day 2005 Commendation, and DRDO Performance Excellence Award 2009.

### Provisioning Hadoop’s Map Reduce in Cloud for Effective Storage as a Service

***Dr.S.Mercy Shalinie - Associate Professor & Head of CSE, Thiagarajar College of Engineering, Madurai***

#### Abstract

Cloud Data Storage has experienced unprecedented growth in recent years. Security concerns in cloud are more prevalent and there is a critical need to secure huge data at rest. For such scenario it is identified that cloud storage can be secured through proper encryption mechanism. Encryption of large dataset being a time consuming process is made easy by performing it through Hadoop’s MapReduce process which executes the operation at faster rate especially for large datasets. Our results prove that encryption followed by compression gives efficient results when the Mapper does the whole job and the reducer writes the content on HDFS. Our experimental proof suggests that MapReduce framework assist in providing a secure storage as a service for the emerging cloud era.

#### Speaker Bio

### Dr. S.Mercy Shalinie is working as Associate Professor and Head of the Department of Computer Science and Engineering, Thiagarajar College of Engineering, Madurai, TN. She received her B.E. Degree in Electronics and Instrumentation in 1989, M.E. in Applied Electronics in 1991 and Ph.D. in Computer Science and Engineering in 2000. She has published 21 papers in referred journals and 45 papers in National and International Conferences.

### She is presently investigating Research and Development projects for DST, DIT and NTRO. Her area of research interest includes Cloud Security and Machine learning paradigms.

### Framework for a suite of algorithms for predictive modeling on Hadoop

***Vaijanath Rao, Rohini Uppuluri - AOL India***

#### Abstract

One of the major challenges in data mining today is predictive modeling from large scale dyadic data involving two sets of entities. Co-clustering methods have been proposed in literature capable of simultaneously clustering the dyadic data by exploiting the duality between the two sets of entities for effective predictive modeling. Some of the example applications modeling dyadic data include ad targeting, personalization of online content etc. Given the large scale of data that needs to be processed, there is a greater need for a robust large scalable framework.

Hadoop provides a platform for large scale data mining and machine learning in a distributed environment with an easy to use distributed programming **Abstract**ion. In this talk, we describe our framework for a suite of some of the existing co-clustering algorithms for predictive modeling of dyadic data on Hadoop. We discuss the map reduce implementation of some of the co-clustering algorithms. We will also show an example application of modeling dyadic data using our framework.

Key Take-away

* Framework covering some of the existing co-clustering algorithms in map reduce
* Data mining at large scale
* Example application of modeling dyadic data using co-clustering algorithms

#### Speaker Bio

Vaijanath Rao is a Technical Lead at AOL India currently leading the efforts on MapQuest search from India. His interests include Machine Learning, Search and large scale data mining. He earned his Masters degree from IIT Bombay in 2005. LinkedIn: http://in.linkedin.com/in/vaijanathrao

Rohini Uppuluri graduated from IIIT Hyderabad with a Masters degree in 2007. She is currently working for AOL India as Software Engineer. Her interests include Machine Learning, Natural language processing. Her current work includes data mining from large scale data and building personalization and recommendation systems.

## Panel Discussion