ABSTRACT. A taxonomy of enterprise data growth is presented. The taxonomy maps data growth in two dimensions, a horizontal axis that classifies data activities – data creation, use and persistence – and a vertical axis that classifies types of data use. Subcategory classifications are presented with examples. An outline of continuing research is discussed.
1.0 INTRODUCTION

Most of the world’s enterprise data was created in the past decade, and the rate of growth of that data is doubling every two years. The drivers are machine growth as well as changes in user behavior. First, the sheer growth in the number of computing devices used by people and organizations continues unabated. It is estimated that over 15 billion devices will be connected to the Internet globally by the end of 2015.1 Equally important are behavioral changes in how people use computers, and why they use them. A decade ago, there were 677 million global internet users, mostly on desktop computers, mostly at home. The International Telecommunications Union (ITU) estimates there are now over 2.5 billion users - Facebook alone is estimated to have 835 million - using a multitude of different kinds of computers, most of them mobile.2 With the year-over-year increases in user populations, service volumes grow at compound rates. Apple iTunes recorded its 25 billionth download in March of 2012, having surpassed 6 billion downloads by early 2009.3 Twitter, established in 2006, is estimated to have over 200 million users tweeting 230 million times a day.4

Dramatic changes in consumer use are also mirrored in the corporate use of technology. Cisco estimates that global internet traffic will quadruple from 2011 to 2016, reaching 90.9 exabytes a month, a compound annual growth rate of 31%.5 A recent study of enterprise servers found that the world’s installed base of servers processed 9.57 zettabytes of data annually, an average of 63 terabytes of data a year for every registered company on the globe.6

The impetus for enterprise data growth is innovation, productivity gains and operating efficiencies. Data itself is becoming the major source of competitive advantage. Developing a capability to collect and refine data into greater business knowledge and increased enterprise productivity are crucial ways that companies can derive value from their data. But how can companies successfully leverage the strategic value of data growth?

In this first of three working papers on data growth and enterprise productivity, we develop a preliminary taxonomy of enterprise data growth. It is preliminary in that each of the three papers addresses a required step in developing, validating and refining a new classification scheme. This paper presents the concepts behind the taxonomy. We first outline a number of important characteristics of data and data growth, and then develop a preliminary index for measuring enterprise data growth. The index will form the basis for a later paper validating the growth taxonomy. We define our data taxonomy as a two-dimensional framework that combines ideas from industry practice, including data creation, data use and data persistence, and concepts of specialization and generalization of use derived from inheritance models in software programming. Our taxonomy maps data growth into different data activities over time, and locates these activities in a hierarchy of specialized and generalized use cases. We conclude with implications and a brief discussion of our second working paper.
2.0 OBSERVATIONS ON DATA

2.1 Defining Data

Data are collections of numbers, characters, images or other outputs from digital devices that represent physical quantities as artificial signals intended to convey meaning. ‘Artificial’ because data is created by machines, such as microphones, cameras, environmental sensors, barcode readers or nanometers, and by people interacting with computers, using keyboards, voice commands (Apple’s Siri), or other input devices. Streams of data from sensors are extensively transformed by a series of machines, such as cable routers (location change), storage devices (time shift), and computers (symbol and meaning change). These transformations, in turn, create new data. Digital data has the desirable properties that it is easy to capture, create, communicate and store: so easy in fact, that we are generating enormous amounts of it. Data is the lowest level of abstraction from which information and knowledge are derived. Information has the further property that it must have meaning for its intended use. People define that meaning, whether it is the information required for an immediate decision or the collection of background information for an action to be taken in the future. The amount of human involvement increases as we move from a focus on data to one of information.

2.2 Drivers of Enterprise Data Growth

At its core, the key driver to enterprise data growth is the value of the data and its assessment through analytics. Data value is defined by use, now or in the future. Three growing market opportunities – social, mobile and cloud – have focused current investor and enterprise attention on data analytics and “big data” – a subjective term referring to datasets whose size and rate of processing at scale is beyond the ability of current database software tools to capture, manage, store and analyze. Data growth itself is the larger and more pervasive enterprise challenge than “big data,” since it encompasses all of the data the enterprise creates, processes and stores for all corporate functions and responsibilities. While “big data” may refer to the knife-edge of the current market focus on analytics, enterprise IT must provision for all of the data an enterprise needs to function, in addition to the “big data” segment.

As part of our research on data growth, we have conducted a series of interviews with Chief Information and Chief Technology Officers in a sample of 30 large enterprises. Our interviews asked about data growth, its drivers and implications, and what steps companies were taking to leverage growth opportunities. In every organization studied, data growth far exceeded revenue growth. The yearly data increases ranged from 18% for a manufacturing company to over 150% for a hospital group. The median compound annual data growth rate was 40% per year, or a rough doubling of the amount of data stored every two years. The highest growth rates were reported in research organizations (where some research activities produce huge data stores) and hospital systems (where radiology and other image-intensive functions and electronic medical record applications drive storage demand).
Below we have abstracted three sets of growth drivers from our interviews and continuing research, in categories of general business, technology and data technology. All were seen by our interviewees as important, with some critical in specific industry segments.

Business drivers include:

- Market leadership: customer quality, market intelligence, best in class
- Business performance: speed, agility, scale, complexity
- Financial performance: investment, cost performance
- Regulation and Compliance: Mandated retention of business records, now and future
- Discovery: Legal discovery in regulated industries
- Information/Analytics: need for real-time, predictive analytics; trend analysis at scale

Technology Drivers:

- Digitalization: data is increasingly everywhere and in many formats
- Moore's Law: machine capacity and performance curves
- New software, hardware architectures: Hadoop, cloud, mobile
- Analytics: Real-time, predictive, visualization
- Machine intelligence: M2M not limited to human thresholds (time, attention, cognitive, etc.)

Data Drivers:

- Data Transparency: merging of private and public data formats (Google Fusion, etc.)
- Data Granularity: technology improvements in digital data devices, sensors, etc.
- Data Engineering: Improvements in data storage software, deduplication, replication, archiving, etc.

Several market initiatives and new companies point to important future drivers in data growth. These include Wolfram’s Computable Data Initiative, Google’s Fusion Tables project, and new start-ups in the large-scale data aggregation space, such as Factual, a company developing an open data platform for application developers leveraging public and private data aggregation and community exchange. Wolfram’s Computable Data Initiative is built around the proposition that as systematic knowledge and data accumulate, the task is somehow to make it all computable. Wolfram notes that structurally getting the data into an analytics software program is the easy part (5% of the effort), and that the hard part is getting the data connected to other things—to other data, to other computations—and then to humans, in order to make results maximally efficient for people to absorb. The data initiatives at Wolfram, Google and Factual all share the common thread of assembling very large collections of systematic data matched with new computational algorithms that extend beyond current modeling methods. They reinforce the general point that the total volume of data will continue to go up, correlated
with the work necessary to make raw data computable, and new algorithms to process the data for human and machine consumption.

2.3 Measuring Data Growth

The measurement of data growth in firms, of course, is not a single number. It is a composite of the growth in data stored in enterprise storage systems ("data at rest"), and data in motion, or "kinetic data," data that is used and generated in computation, and in flow over enterprise networks. Digital data is also created by sensor systems and other smart digital devices. The great majority of data created in enterprises today is unstructured, made up of all the messaging, email communications and office documents produced by mobile and office applications, including image, audio and video files that are increasingly important in social media applications. Structured data makes up the rest of enterprise data. As companies invest in enterprise resource planning systems (ERP), customer relationship management systems, RFID, sensors and other device technologies, the amount of data associated with transactions multiplies. Companies collect data on inventory levels, transportation movements, and financial transactions, enabling them to communicate more accurately with customers, optimize supply chains and business processes, and manage financial risks. Ideally, transaction data is collected and stored once. In practice, many organizations have redundant applications and databases, which add to data storage costs and can make data more difficult (and more costly) to access.

2.4 An Extreme Growth Example

How the different constituent parts of large scale data grow relative to one another is an important question (that is, does the growth of each part – storage, computation, and network transport - maintain a constant scaling factor over time). There is evidence to suggest that rapid technological progress in science-based fields may be radically changing growth ratios. According to Kahn (2011), the field of genomics is increasingly dominated by the growth in the size of data and the effective use of processed (derived) information. Output from next-generation sequencing machines has grown in output from approximately 10MB per day to 40 GB per day on a single sequencer, with the total number of sequencers worldwide approaching 200. Kahn states that this growth in raw data output has outstripped Moore’s Law advances in computational speed and storage capacity in laboratory information systems, forcing reexamination of the definition and value of “raw” and “derived” information, and affecting progress on developing community data repositories (shared data exchanges) in the years ahead (Fig. 1). A related challenge is analyzing all of the data effectively - the pace of innovation in data creation is currently much higher than the rate of innovation in genomic informatics. Kahn states that this widening gap must be addressed for genomic research to take advantage of the enormous growth in raw sequencing data.
For our purposes, extreme data growth in genomics and the acute challenges posed both to researchers and R&D laboratories leaves open the question of how this case generalizes to other use cases in science-based and business organizations. It is certainly suggestive, but it is an empirical question as to how many other information- and analytics-driven organizations may be similarly affected. In the next section we develop a prototype enterprise data index that we will use to collect and analyze field data in business organizations to help answer this question.

2.5 A Prototype Enterprise Data Growth Index

Our goal is to estimate the relative growth of the different components of enterprise data. To do this, we define a prototype data growth index as follows: the sum of the average daily online storage available (in bytes), plus the average daily network throughput rate in bytes/sec \( X \) the number of datacenter network switch ports, plus the average daily server transaction rate in transactions/sec \( X \) the number of datacenter transaction servers, normalized for the different time units (Fig. 2).
To illustrate how the index works, consider the following example. A hypothetical company has an average daily online storage capacity of 100TB. The company’s 1,000 network switch ports maintain an average network throughput rate of 100 MB/sec over a 24 hour period. Assume further that the firm’s 2,500 Web and database transaction servers process, on average, 1,000 transactions/sec daily (assume each transaction is 8K bytes). Based on these assumptions, we calculate that the total average daily volume of kinetic and stored data activity in the firm is just over 1 PB, on an online data storage capacity of 100 TB. This gives a ratio of approximately 10 to 1 comparing “kinetic” data – data generated in computation and in flow over the enterprise network, to data at rest.

This index, of course, is highly simplified. The operating environments of each of these IT components is highly context-specific and subject to large variance in daily operations, workload performance, and utilization rates. However, an important goal in developing the index is to include it in questions we are asking datacenter managers and other IT personal in field surveys and case analyses of data growth. As such, the abstraction level of the index is that which a datacenter manager routinely collects or can estimate with accuracy. As fieldwork progresses, we will refine the index as appropriate.

### 3.0 DEFINING AN ENTERPRISE DATA GROWTH TAXONOMY

In this section we present a taxonomy of enterprise data growth (Fig. 3). The goal of the taxonomy is to develop a classification that is simple enough to be empirically tested and sufficiently comprehensive to include all of the major factors in data growth. A summary of the taxonomy is presented below, with the important elements discussed in the following sections.
The taxonomy maps data growth in two dimensions, a horizontal axis that classifies consecutively the major “activities” of data – data creation, data use, and data persistence – and a vertical dimension that classifies types of data use. Each of the main data activities can be broken down further into more detailed sub-activities. For example, the ways in which data is created, by machines or by humans interacting with machines, or more specialized types of data, for example, data created by sensors, a type of observational data. The vertical axis represents specializations or, alternatively, generalizations of data use. For example, one category of data use is transactional, and one type of transaction is the online purchase of a product or service. When this type of transaction is processed, data is created to record, process the purchase, and store the transaction record. The generalization of this transaction is to include different activities of which this one is a type. For example, customer activities prior to the purchase - searching online for the product, browsing substitute products, comparing prices and so on, are generalizations of the purchase transaction.

Using this approach, classifying different data use cases can be located Down the vertical axis in the taxonomy, corresponding to different types of use (specializations), or Up the vertical axis, corresponding to aggregate classes of use, of which the one use is a type (its generalizations).

Details of the important elements of the taxonomy are discussed below – sub-categories within data creation, data use and data persistence. Example data types, work activities and other attributes are discussed.
3.1 Data Creation

Digital data creation is driven by the uses of data, the capture of data through sensors and other data sensing and recording technologies, and the replication, duplication and reproduction of data in multiple formats. In a sense all operations on digital data produce new data, and that data is either overwritten moments later, or retained over a period of time.

Machines and people interacting with machines create digital data. We have classified five categories of data creation: observational, transactional, authored, simulated, and in process (Fig. 4). For example, people interacting with computers “author” digital data, in the form of documents, spreadsheets, emails, text messages and so forth. The volume of authored data can be approximated by taking into account the number of users, the time taken in authoring, and the capacity of the device used to author. Another type of data creation is “Observational”, that is, data generated by sensors and other digital recording devices (machines). Where people initiate the recording activity, they do so with devices such as digital cameras (taking a digital picture) or smartphones (making a video recording). Note that if a user takes a picture (observational data), and then edits that picture on a notebook computer, the output from the editing session produces another type of data, in this circumstance authored data. Data creation is not restricted to any one type of data created in any one defined sequence.

<table>
<thead>
<tr>
<th>Data Creation</th>
<th>Human / Machine Generated</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observational</td>
<td>Human, Machine Generated</td>
<td>Human-initiated: Data capture through smartphone cameras, digital video recorders, other recording devices. Bound by the number of devices in use; average time of use; etc. Machine-generated: # sensors deployed, installed base of digital machines, M2M. Bound by # of sensors in use, sensor bandwidth.</td>
</tr>
<tr>
<td>Transactional</td>
<td>Human, Machine Generated</td>
<td>Human-initiated: exchange or transfer of goods, services, or funds; average transactions/human. Machine-generated: M2M; average machine transaction rates/bandwidth.</td>
</tr>
<tr>
<td>Authored</td>
<td>Human</td>
<td>People interacting with computers and personal digital devices to author digital content; documents, emails, spreadsheets, IM, tweets. Bound by the number of users, the average rate of authoring, the capacity of the digital device in use.</td>
</tr>
<tr>
<td>Simulated</td>
<td>Machine</td>
<td>Machine generated. Data created for simulation modeling. Rate tied to the number, growth rate of simulation algorithms/applications.</td>
</tr>
<tr>
<td>In-Process</td>
<td>Machine</td>
<td>Machine generated. The number of applications that process, store and transmit data (tied to the number of humans who use the applications); applications that replicate, copy, de-duplicate, or otherwise affect data stocks; rate of data processing and intermediate result generation, storage and transmission of results.</td>
</tr>
</tbody>
</table>

Fig. 4. Data Creation
Simulated data is an important type of machine generated data. This is data generated by algorithms and/or applications for the purposes of representation and analysis. Simulated data is used extensively in science and industry, in cases where the primary collection of data is not technically or economically feasible, or where the analysis of simulated data can be used to study unpredictable events, such as the path of a tornado or the failure of a power station in an electrical grid due to a lighting strike or other natural calamity.

### 3. 2 Data Uses

Data uses can be defined narrowly, to mean for example, data that is used in specific information systems, such as customer relationship management (CRM) or enterprise resource planning (ERP), or in business functional areas such as manufacturing, product engineering, human resources or sales. It can also be defined broadly, to include the purpose or goals of data use – to inform (informational), to interact (communications), to complete work (work production), or to pursue scientific discovery. For our purposes, we have subdivided data uses into broad use categories. We will capture narrower definitions of use as specializations in the taxonomy, as explained earlier.

The five use categories are (Fig. 5):

- Knowledge Production, which incorporates the range of activities taken to transform data into knowledge
- Interaction, which includes communications, transaction activity, and social interaction
- Work Production, which includes work task and work process outputs, and task/process automation
- Regulation, the social, legal and governmental requirements for information disclosure and compliance
- Scientific Research and Discovery, the progress of current science and future science
<table>
<thead>
<tr>
<th>Data Use</th>
<th>Use Attribute</th>
<th>Example Use Objectives, Action</th>
<th>Example Decision, Work Processes, Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Knowledge Production</td>
<td>Informational</td>
<td>Inform people, machines, processes</td>
<td>Conduct market intelligence on a new product or service.</td>
</tr>
<tr>
<td>Analytical</td>
<td></td>
<td>Apply complex algorithms to computable data</td>
<td>Analyze seismic field data (image data). Oil &amp; gas industry.</td>
</tr>
<tr>
<td>Knowledge creation</td>
<td></td>
<td>Assimilate, systematize information into user knowledge</td>
<td>Therapeutics, biological research and development. Pharmaceutical industry.</td>
</tr>
<tr>
<td>Interaction</td>
<td>Communication</td>
<td>All people and machine communications</td>
<td>Email, text messaging, video conferencing.</td>
</tr>
<tr>
<td>Transaction</td>
<td></td>
<td>Communicative action; completion of exchange or transfer of goods, service or funds</td>
<td>Issuing an invoice and collecting payment online.</td>
</tr>
<tr>
<td>Social</td>
<td></td>
<td>Develop, discover relationships, relational structure in social network data at scale</td>
<td>Post, share information on a social media website.</td>
</tr>
<tr>
<td>Work Production</td>
<td>Work task, work process output</td>
<td>Design, engineering of physical and/or informational products and services; production and distribution of services</td>
<td>Design and build a new product, service.</td>
</tr>
<tr>
<td>Process automation, Redesign</td>
<td></td>
<td>Informate, automate business processes; augment, redistribute, reduce or replace human labor</td>
<td>Supply chain automation; automated customer service.</td>
</tr>
<tr>
<td>Regulation</td>
<td>Business regulation, legal compliance</td>
<td>Financial regulation (SEC), product safety regulation, healthcare information privacy (HIPPA)</td>
<td>Voluntary and mandated financial disclosures, internal audit and compliance, corporate legal</td>
</tr>
<tr>
<td>Scientific Research and Discovery</td>
<td>Knowledge discovery</td>
<td>Create, instrument new data and information collection, algorithms and analytic methods.</td>
<td>New fields of computational linguistics, genomics, and computational astrophysics</td>
</tr>
<tr>
<td>Future Science</td>
<td></td>
<td>New fields of investigation generated by new systemization of knowledge; data creation more amenable to automation (computable data+algorithms)</td>
<td>Computable data examples: Computational Finance, Linguistic Data, Geospatial Data</td>
</tr>
</tbody>
</table>
For example, in Knowledge Production, data is used as input in computation and analysis conducted to inform people, or other machines, of a specific result. A related subcategory is Analytical, the application of models or algorithms to data for purposes of analyzing that data.

The categories Interaction and Work Production encompass most of what people do on an everyday basis in an enterprise – they interact socially with other people and functionally with machines, with the goals of communicating and completing work. Increasingly the instruments used for both activities are digital or digitally mediated. In our office environments, who uses pen and paper anymore? Or relies on voice phone calls for all communications? These and most other office functions are now digital, reliant on digital data.

At this level, our data use categories are necessarily abstract, in order to capture in as comprehensive a way possible what is meant by use. As our research work progresses, an important objective will be to elaborate these use categories and their specializations through case examples. Included in this objective is the task of researching and analyzing new use cases. Note also that several important categories of data use are not included in this list, for example, leisure and entertainment. People playing computer games, watching high-definition television, or watching video on their tablet computers consume a lot of digital data. We do not include these use categories here as they are specific to consumptive (media) use by consumers. Of course, enterprises produce, deliver and manage the great majority of media data presented to consumers for entertainment consumption. In enterprises however, we are concerned with the data activities for the completion of productive work.

### 3.3 Data Persistence

Data persistence encompasses two basic data storage questions faced by all businesses: what data (or data types) do we store, and for how long do we keep it? In practice, data persistence refers to the time- and policy-based rules for migrating data across different performance and cost tiers in an enterprise storage system, as that data ages or is judged to be of lesser business value. Regulation drives many of these decisions – HIPAA for example - and in many cases government mandated data retention periods exceed customary technology replacement cycles. In these cases, data ultimately may require migration from one generation of storage system to another. Large scale data migration in these circumstances can easily cause a myriad of technical and policy problems.

Data persistence also has a broader meaning that includes greater reference to management practice in defining data policies. If we consider data creation to be primarily about discovery, we can think of data persistence as primarily about the propensity to extend time scales in storing valued things. That value, of course, can be highly subjective over time, especially when management changes occur. Software applications such as business analytics, data mining, and decision support systems provide the ability to sustain value from historical data over time. Deployment, therefore, can have the effect of further lengthening data time scales.
<table>
<thead>
<tr>
<th>Data Persistence</th>
<th>Example Data Characteristics</th>
<th>Example Work, Process Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temporal</td>
<td>In-process data created and stored briefly in processor, network and storage computation, input-output operations and overhead; old data is overwritten, stored only for fractions of a second as new data arrives.</td>
<td>Time and data-intensive business activities within functional areas. Business analytics. Examples: financial trading systems, oil &amp; gas field seismic analysis, fraud detection in credit card processing and risk analysis.</td>
</tr>
<tr>
<td>Active</td>
<td>The active data stores of an organization, typically stored on a tiered storage system with multiple tiers defined by frequency of data access and data use (application) attributes. “Primary” data.</td>
<td>All core business process IT systems: CRM, ERP, financial analysis and reporting systems, customer acquisition, supply chain management, product engineering and manufacturing systems.</td>
</tr>
<tr>
<td>Retained</td>
<td>Data that is backed-up, copied, replicated, de-duplicated and/or otherwise retained for purposes of distributed access, performance, data security and integrity, and disaster recovery.</td>
<td>IT: Enterprise data management systems and software. Backup and disaster recovery. Data protection, data access and security.</td>
</tr>
<tr>
<td>Historical</td>
<td>Aged active data that has been moved to lower performance, lower cost storage tiers. Sometimes referred to as secondary or tertiary data.</td>
<td>Industry practice varies widely on data movement across storage tiers. Typical factors involve frequency of data access, data use / application domain, response time, I/O performance and storage cost.</td>
</tr>
<tr>
<td>Archive</td>
<td>Fixed content, regulatory compliance and archive data that has moderate to low activity, and long-term if not infinite retention requirements. The retention time scale can be the lifetime of the organization, or longer.</td>
<td>Examples: Healthcare records (HIPPA), financial and insurance data, legal records, pension fund member data, customer records. Email record archives.</td>
</tr>
</tbody>
</table>
We have subdivided Data Persistence into five time-based categories: Temporal, Active, Retained, Historical, and Archived (Fig. 6). Temporal data refers to data created in process – machine data that is generated in computation (for example, in data analytics) but is not retained. On the other end of the classification, Historical data is active data that has aged, and likely has been migrated to lower performance, lower cost storage tiers (tier 2 or tier 3 storage). The decision to migrate Historical data to lower cost tiers may be technically driven (based on rule-based storage management software for example), but usually in the case of higher value data the migration decision is based on a combination of technical criteria and management data policy. What is important to recognize is that data time scales can set very lengthy requirements for policy documentation and access. Archived data in a pension fund management company, for example, may need to be retained for 150 years or more (two generations, the pensioner and the children). In such cases there is little to assure that policies applied to the data when it was migrated will persevere over such a lengthy period.

4.0 DISCUSSION AND FUTURE RESEARCH

Our goal in this paper has been to develop and present a prototype taxonomy for enterprise data growth. The taxonomy is based on two central concepts, a horizontal axis that arranges consecutively the major “activities” of data – creation, use and persistence – and a vertical axis that represents specializations or generalizations of how data is used in the enterprise. The background concepts are drawn from industry practice and software programming. We then divided each data activity into subcategory classifications and provided descriptions and an example for each subcategory. Prior to our taxonomy discussion, we presented our preliminary work on building an enterprise data index. The purpose of the index is to separate out the main components of data growth, and suggest how to measure those components and compare them. Both the data index and taxonomy are prototypes and subject to further refinement.

Our motivation in researching data and data growth are several: first, we appear to be at a critical inflection point in our understanding of how Moore’s Law improvements in compute, network and storage capacities are ushering in new paradigms in data intensive computing. Server compute capacities are doubling every other year, driving similar growth rates in stored data and network capacities. In the face of rapid changes in core infrastructure capabilities, many firms are finding it necessary to completely rethink their approaches to corporate IT for performance, cost and economies of scale, and in view of new industry initiatives in cloud computing, “big data”, data analytics and green datacenters. Secondly, we need more and better use case analyses of how companies are leveraging the opportunities in data growth – where is the value in all of this data? More and better recording and analysis of emerging, successful practices is important.
This working paper is the first of three papers on enterprise data growth. Our next research phase focuses on field studies and use case analyses of data growth. Our goals are to refine and validate the data index and growth taxonomy, and to record, analyze and ultimately share successful industry practice. We understand this will be an iterative process and require the support of research colleagues and industry experts to accomplish our objectives. We invite those interested to contact us and join in.
Acknowledgments

This working paper is the product of industry and university collaboration. We are grateful to our industry sponsors and research partners. We extend special thanks for the contributions and support provided by Larry Freeman and Cathy Gadecki of NetApp, and Jeff Burke of Seagate Technology.

Financial support for the Center for Large Scale Data Systems Research (CLDS) at the San Diego Supercomputer Center is gratefully acknowledged. Our corporate sponsors are:

Brocade  
Greenplum  
NetApp  
Seagate Technology

Additional support has been provided by:

Mellanox  
The National Science Foundation

The authors bear sole responsibility for the contents of this working paper.

Questions about the paper may be addressed by contacting the authors:

Jim Short, jshort@ucsd.edu  
Chaitan Baru, cbaru@sdsc.edu

Center Website: http://clds.ucsd.edu
References and Notes

2 International Telecommunications Union (ITU) data. Source: http://www.itu.int/ITU-D/ict/
4 Twitter does not publish usage data publicly. This data is taken from a September 2011 news conference with CEO Richard Costolo. Source: http://www.huffingtonpost.com/2011/09/08/twitter-stats_n_954121.html
14 For further information on index construction and sample calculations, please contact the authors.