Process Improvement in Cybersecurity and Big Data

Angel Jordan
University Professor Emeritus
Provost Emeritus
Carnegie Mellon University

CIMPS 2014

OCTOBER, 2014

Summary

This presentation begins with a quick status of conventional capability and maturity models for Process Improvement in software engineering.

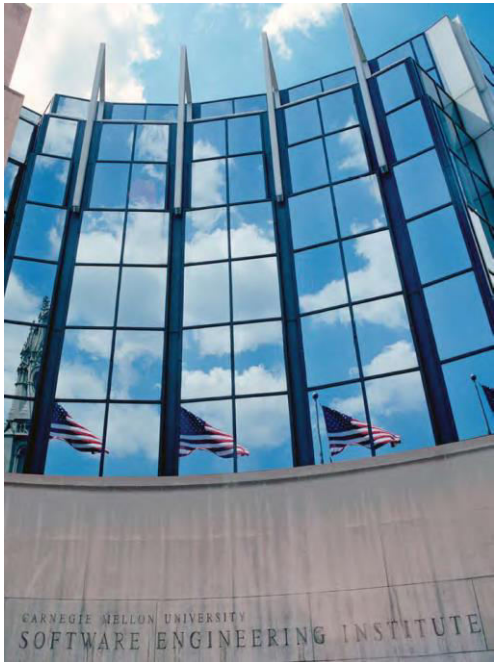
It then deals with Process Improvement in Cybersecurity with emphasis in the CERT-Resilience Management Model.

Recent developments in Big Data are addressed and the Link between Big Data and Process Improvement is introduced.

The presentation culminates with recent initiatives involving Big Data at Carnegie Mellon University and the Software Engineering Institute.

The Software Engineering Institute

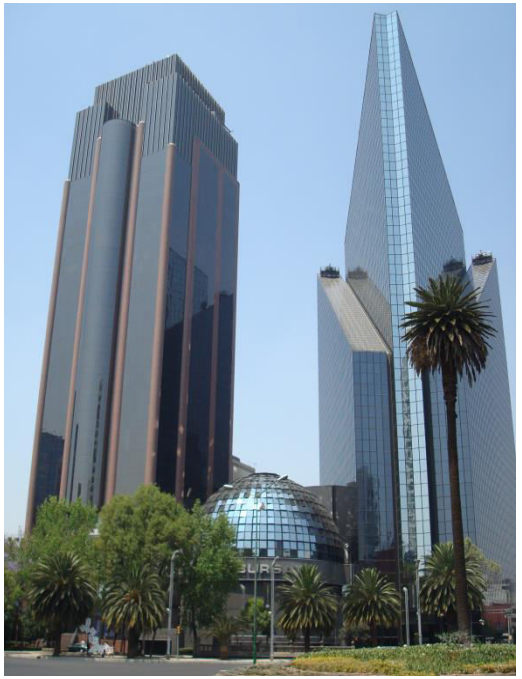
- Department of Defense R&D Laboratory; FFRDC
- Operated and managed by Carnegie Mellon University
- Created in 1984
- Headquartered in Pittsburgh, PA
- Chartered to provide support worldwide



A Broad Spectrum of Stakeholders

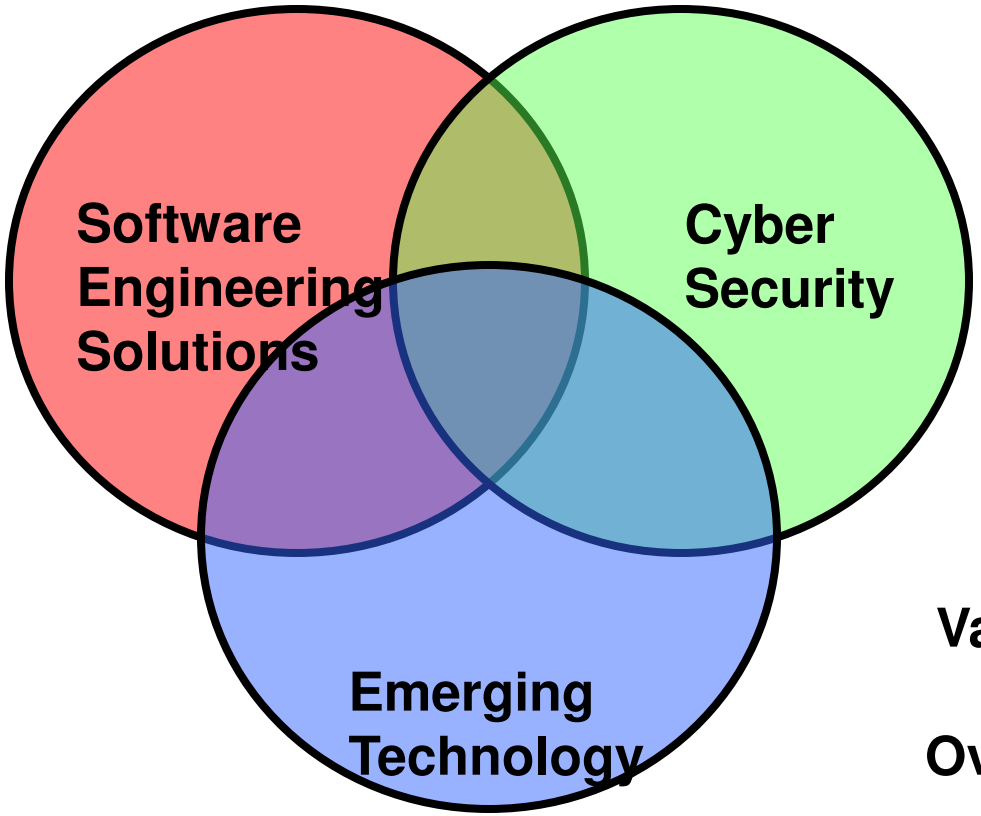
The SEI advances research in software engineering and cyber technologies for its stakeholders:

- Major government customers
  - U.S. Department of Defense (DoD)
  - U.S. Department of Homeland Security (DHS)
- Key industries and organizations with the potential to advance software engineering and related disciplines
- Researchers, developers, users, and acquirers — government, commercial, and academic
- Strategic Partners worldwide



Bolsa Mexicana de Valores (BMV)

SEI Technical Program Structure



SEI's Special Value Proposition is in the Overlapping Areas

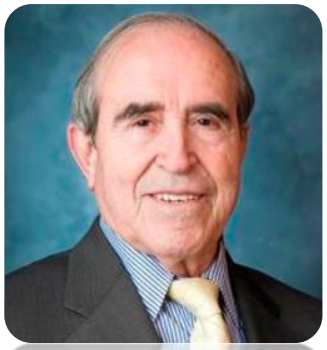
What happened to Process Improvement?

Thirty Years of Process Improvement and Counting
CMM, CMMI, and Process Management

The CMMI Institute.
http://whatis.cmmiinstitute.com/

Management of the Software Development Process

- DoD *Software Technology for Adaptable, Reliable Systems* (STARS) strategy developed (1982)



- SEI founded in 1984
- SEI strategic plan—supported by DoD and defense contractors—recognized as a fundamental activity (Barbacci 1985)
- IBM had mature efforts that had proven effective (Humphrey 1985)

- The SEI recruited Watts Humphrey from IBM
  - Process Management Framework Project in 1986

- The Air Force Program Manager asked the SEI to conduct a study of “best practices.”
  - SEI used this customer interest to drive the Process Management Framework project

Expansion of Maturity Modeling

After the first CMM for use in software development, we asked:

"Why not create a powerful framework for processes in multiple disciplines?"

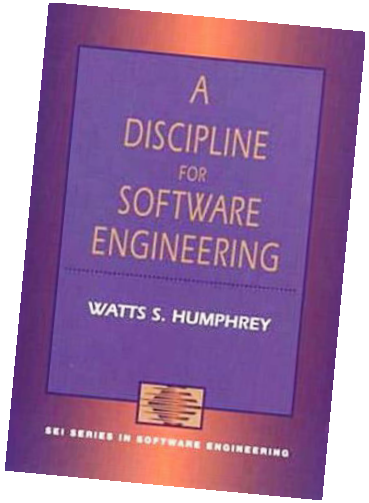
The SEI began developing maturity models for other areas:

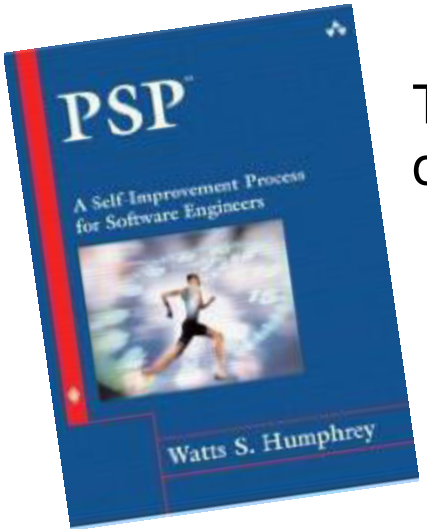
- People CMM, for managing human assets
- Systems Engineering CMM
- CERT Resilience Management Model

Bringing Discipline to Software Development

If CMM had an impact on the management system:

- Why not extend it to the people who actually do the work—the practicing engineers? (PSP TR, Humphrey 95)
- The Personal Software Process (PSP) was built on the principle that every engineer who works on a software system must do quality work to produce a quality system.





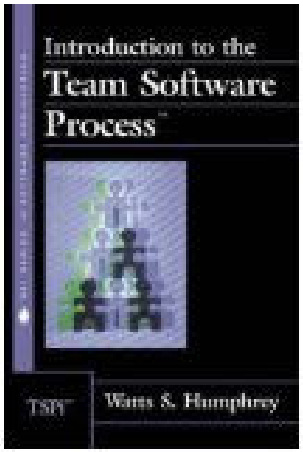
The development of the PSP began with the application of CMM principles to writing small software programs.

- Further refinement led to a disciplined personal framework.
- It allows engineers to establish and commit to effective engineering and management practices for their software projects.

Effective Engineers Must Be Able to Work on Teams

The Team Software Process (TSP) was developed to help development teams establish a mature and disciplined engineering practice.

- **Goal**: produce secure, reliable software in less time and at lower costs.



The TSP has been applied in small and large organizations in a variety of domains with documented results:

- productivity improvements of 25% or more
- higher quality software with fewer defects
- testing costs and schedule reductions of up to 80%
- cost savings of 25-50% per software product release

In the Late 1990's, DoD Faced a Set of Problems

- Risks and threats from technology advances were present in organizations in every sector
  - US federal government agencies, defense and commercial industry, and academia
- The SEI recognized that the best practices of such organizational challenges could best be managed with a capability maturity model.
- Over the ensuing ten years, the SEI engaged the relevant communities in evolving such a model.

This work culminated in the CERT Resilience Management Model, released in 2010.

SEI toward Security & Survivability

The **SEI** is home to **CERT**, which studies **internet security in networked systems,** and develops information and **training to help you improve security.**

Application of technologies and processes to **achieve** a **required** level of **confidence** that software systems and services **function in the intended manner**, are **free from accidental or intentional vulnerabilities**, provide **security capabilities** appropriate to the **threat environment**, and **recover** from **intrusions** and **failures**.

Computer Emergency Response Team

The name **computer emergency response team** is the historic designation for the first team (CERT/CC) at Carnegie Mellon University (CMU). **CERT** is now a **registered service mark of Carnegie Mellon University** that is **licensed to other teams around the world**.

Some teams took on the more generic name of **CSIRT (Computer Security Incident Response Team)** to point out the task of **handling computer security incidents** instead of other tech support work. Because **CERT is a registered trademark owned by Carnegie Mellon University, it should not be used interchangeably with CSIRT.**

The name CSIRT
Computer Security Incidence Response Team

The history of CSIRTs is linked to the existence of malware, especially computer worms and viruses. Whenever a new technology arrives, its misuse is not long in following. The first worm in the IBM VNET was covered up. Shortly after, a worm hit the Internet on 3 November 1988, when the so-called Morris Worm paralyzed a good percentage of it. This led to the formation of the CERT/CC at Carnegie Mellon University under a U.S. Government contract. With the massive growth in the use of information and communications technologies over the subsequent years, the now-generic term "CSIRT" refers to an essential part of most large organizations' structures.

CERT Founders:

Druffel, Pethia, Scherlis, Squires

http://www.us-cert.gov/

United States Computer Emergency Readiness Team

http://itlaw.wikia.com/wiki/US-CERT

The **United States Computer Emergency Readiness Team** (**US-CERT**) has played an important role in public sector data security

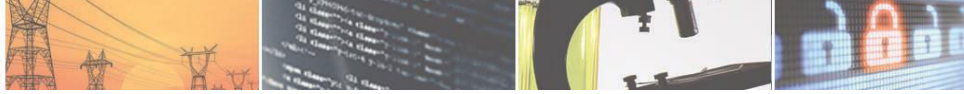
US-CERT is a partnership between the Department of Homeland Security (DHS) and the public and private sectors

It is currently positioned within the National Cyber Security Division (NCSD) of DHS's Office of Cybersecurity and Communications

Established in 2003 to protect the nation's Internet infrastructure, US-CERT coordinates the nation's efforts to prepare for, prevent, and respond to cyber threats to systems and communication networks.

The organization interacts with federal agencies, state and local governments, industry professionals, and others to improve information sharing and incident response coordination and to reduce cyber threats and vulnerabilities

It serve as a focal point for the government's interaction with federal and non-federal entities on a 24-hour-a-day, 7-day-a-week basis.

CERT Focus Areas

Strengthen Foundations

Science of Cybersecurity

Develop research methods, system and human performance metrics, and simulated test environments, to research when and why systems become unsecure

Acquire and Develop

Secure Coding

Produces standards, techniques, and tools that software developers and software development organizations require to eliminate vulnerabilities before software is deployed

Cybersecurity Engineering

Research needed to prepare acquirers, developers, and operators of large-scale, complex, networked systems to improve security, survivability, and software assurance

Operate and Sustain

Resilience Measurement and Analysis

Creates qualitative and quantitative frameworks and models for assessing, measuring, and improving security and resilience in complex systems and critical infrastructures

Insider Threat Modeling and Analysis

Performs research, modeling, analysis, and outreach to define socio-technical best practices so organizations are better able to deter, detect, and respond to evolving insider threats.

Monitor, Detect, Investigate, and Recover

Cyber and Analysis

Creates and applies novel techniques, tactics and procedures that help the DoD effectively conduct full spectrum cyber operations on systems and networks critical to national, homeland, and economic security

Digital Intelligence and Investigation

Focuses on large scale incident response techniques and methodologies; tools and techniques for rapid data triaging; low level hardware forensics; and the ongoing analysis of disruptive technologies and emerging threats.

Cyber Workforce Development:
Develop methods, tools, and metrics to develop, train, and exercise a mission-ready cyber workforce

Software Assurance

Software assurance (SwA) is defined as "the level of **confidence** that software is free from vulnerabilities, either **intentionally designed** into the software or **accidentally inserted** at anytime **during its lifecycle**, and that the **software functions in the intended manner."**

QA (Quality Assurance) is also applied to software to verify that **features and functionality meet business objectives,** and that **code is relatively bug free** prior to shipping or releasing new software products and versions.

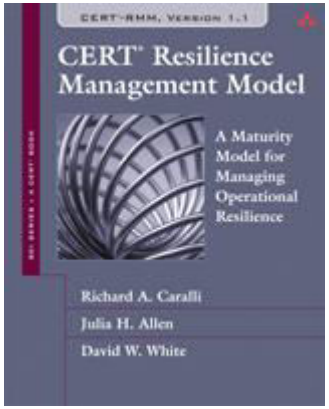
Assurance in QA a companion of Testing. Going back to 2004 in China before the boom of Chinese software.

CERT Resilience Management Model

A maturity model that promotes the convergence of security, business continuity, and IT operations activities to help organizations manage operational resilience and risk

CERT-RMM V1.0 is available as a free download.

Version 1.1 of the CERT-RMM was published in a book by Addison-Wesley Professional in December 2010.



CERT-RMM Capability Appraisals

Capability appraisals are an objective way to determine your organization's current level of capability for managing operational resilience based on the model's capability level scale.

CERT-RMM Training

History of CERT-RMM http://www.cert.org/resilience/products-services/cert-rmm/history.cfm

CERT Resilience Management Model

CERT-RMM has two primary objectives:

- Establish the convergence of operational risk and resilience management activities such as security, business continuity, and aspects of IT operations management into a single model.
- Apply a process improvement approach to operational resilience management through the definition and application of a capability-level scale that expresses increasing levels of process improvement.

The model combines aspects of IT operations management with operational risk and resilience management, such as security and business continuity.

The model does the following:

Provides a process definition, expressed in more than **20 process areas** across **four categories**: **enterprise management, engineering, operations management, and process management**

Focuses on **four essential operational assets**: **people, information, technology, and facilities.**

Includes processes and practices that define a **scale** of **four capability levels for each process area: Incomplete, Performed, Managed, and Defined**

Serves as a **meta-model** that includes **references to common codes of practice** such as **ISO 27000, ISO 2230, ITIL, CobiT, and SO24762**

Includes **process metrics and measurements** that can be used to ensure that operational resilience processes are performing as intended

Facilitates an objective **measurement of capability levels** via a structured and **repeatable appraisal method**

CERT-RMM

http://www.cert.org/resilience/products-services/cert-rmm/cert-rmm-model.cfm

CERT-RMM doesn't replace an organization's best practices: Rather, **it provides a process structure** into which **they can be inserted and managed.** The organization can then conduct an appraisal to measure whether the implemented practices are providing the expected results.

From the **Download page**, you can download these CERT-RMM materials:

- **CERT-RMM V1.0 process areas**
- **CERT-RMM V1.0 generic goals and practices**
- **CERT-RMM V1.0 glossary**
- **Addendum** to CERT-RMM V1.0 and CERT-RMM V1.1, **Measures for Managing Operational Resilience**

CERT-RMM Training

http://www.cert.org/resilience/products-services/cert-rmm/cert-rmm-training.cfm

The following courses related to CERT-RMM are available as SEI training.

- Introduction to the CERT Resilience Management Model
- CERT Resilience Management Model (CERT-RMM) Users Group Workshop Series

CERT Resilience Management Model Appraisal Boot Camp

CERT-RMM Capability Appraisals

One of the **features** of the CERT Resilience Management Model (CERT-RMM) is the **CERT-RMM capability appraisal for process improvement** (CERT-RMM appraisal).

Designed to objectively **review** an organization against the **benchmark** of the model's **processes and practices**.

Can be used ***internally*** by an organization **to improve its processes** for managing operational resilience, or

Can be applied ***externally*** to determine the **capability of a third-party organization.**

Either way, the appraisal provides a **foundation for long-term process improvement.**

What distinguishes a CERT-RMM appraisal?

Unlike assessments, audits, or evaluations in the security, business continuity, or IT domains, **the CERT-RMM appraisal** is **designed** to help an organization understand its **level of capability** through an **examination of process maturity**.

In other words, the CERT-RMM appraisal determines not only whether an organization is doing the *right things right now*, but whether it is **capable of sustaining an acceptable level of performance during times of stress and over the long run** as risk environments continue to evolve and change.

In contrast, most practice-based **assessments** focus on **how well the organization meets the prescribed practice at a point in time**, which **fails** to tell the organization whether **it can sustain an adequate level of performance** after the assessment is over.

CERT-RMM Training

http://www.cert.org/resilience/products-services/cert-rmm/cert-rmm-training.cfm

The following courses related to CERT-RMM are available as SEI training.

- Introduction to the CERT Resilience Management Model
- CERT Resilience Management Model (CERT-RMM) Users Group Workshop Series
- CERT Resilience Management Model Appraisal Boot Camp

Big Data

Big data collection of data sets **large and complex**… **difficult to process** using on-hand data management tools **or traditional data processing applications.**

The, **challenges: curation, storage, search, sharing, transfer, analysis and visualization.**

The **trend to larger data sets** is **due** to the additional information from analysis of a **single large set of related data-allowing** correlations to **find business trends**

Data sets in **meteorology**, genomics, **physics simulations**, biological and environmental research. **Internet search, finance**.

Big Data 2

Data sets grow in size from:

Information-sensing mobile devices, **aerial sensory technologies**

remote sensing, cameras, microphones, **(RFID) readers**, wireless **sensor networks,**

World's technological per-capita capacity to store information has roughly **doubled every 40 months since the 1980s**

As of 2012, **every day 2.5 exabytes (2.5×10^{18}) of data were created**.

Big data is **difficult** to work with **using relational data base systems and conventional techniques.**

Requiring instead **"massively parallel software** running **on tens, hundreds**, or even **thousands of servers**"

it may take **tens or hundreds of terabytes before data size becomes a significant consideration.**

Terabytes and Big Data Definitions

1 TB = 1000000000000bytes = 10^{12}bytes = 1000 gigabbytes.

Big Data Definitions: What's Yours?

http://www.forbes.com/sites/gilpress/2014/09/03/12-big-data-definitions-whats-yours/

Principles of Big Data Systems: You Can't Manage What You Don't Monitor.

Four Principles of Engineering Scalable, Big Data Software Systems

Addressing the Software Engineering Challenges of Big Data

http://blog.sei.cmu.edu/post.cfm/challenges-big-data-294

Terabytes and Big Data Definitions-2

Big Data Definitions: What's Yours?

Here's how the OED defines big data: (definition #1) "**data** of a **very large size,** typically to the extent that **its manipulation and management present significant logistical challenges."**

(#2) **"an all-encompassing term for any collection of data sets so large and complex that it becomes difficult to process using on-hand data management tools or traditional data processing applications."**

Defining big data as (#3) "**datasets whose size is beyond the ability of typical database software tools** to **capture, store, manage, and analyze,"**

(#4) "**The ability of society to harness information in novel ways** to produce useful insights or goods and services of significant value" and "…**things one can do at a large scale that cannot be done at a smaller one, to extract new insights or create new forms of value."**

Big Data Definitions-3

(#5) “The broad range of new and massive data types **that have appeared over the last decade or so.”**

(#6) The new tools helping us find relevant data and analyze its implications.

(#7) **The convergence of enterprise and consumer IT.**

(#8) **The shift (for enterprises) from processing internal data to mining external data.**

(#9) The shift (for individuals) from consuming data to creating data.

(#10) The merger of Madame Olympe Maxime and Lieutenant Commander Data.

#(11) The belief that the more data you have the more insights and answers will rise automatically from the pool of ones and zeros.

#(12) A new **attitude by businesses**, non-profits, government agencies, and individuals that **combining data from multiple sources could lead to better decisions.**

Principles of Dig Data Systems: You Can't Manage What You Don't Monitor

http://blog.sei.cmu.edu/post.cfm/principles-big-data-systems-monitor-223

The **term big data a subject of much hype** in both government and business today.

The **cause of all existing system problems** and, simultaneously, the **savior that will lead to the innovative solutions** and **business insights of tomorrow.**

All this hype fuels predictions such as:

From IDC the **market for big data will reach $16.1 billion in 2014**, **growing six times faster than the overall information technology market**,

From a software-engineering perspective, the **challenges** of big data are **very clear**, since they are **driven by ever-increasing system scale and complexity.**

The Challenges of Scale

The **requirements of scalability mandate new design and engineering approaches** because many **existing tenets of software engineering simply don't hold at scale.**

There will be an **increasing number of software and hardware failures as the scale of an application increases.**

Component failure must be seen as the norm, and **applications** must be **designed to be resilient to failures**.

Software **failures induced by scale may occur** in application **components** or in **third-party components integrated into the system, both open-source and commercial.**

The Challenges of Scale-2

Second **characteristic** is that, **as scale increases, so does complexity**.

There are **more component interactions**, **unpredictable request loads on data collections**, and **increased competition for shared resources,** including **CMUs on multicore nodes**, **memory, cluster interconnects, and disks.**

This **inherent complexity** makes **diagnosing aberrant behavior** an immense challenge.

If **performance** suddenly **becomes a problem**, it can be **immensely time-consuming and challenging to diagnose**, whether the cause lies in the transaction implementation itself or is a result of unexpected interactions with other components.

The Challenges of Scale-3

A third characteristic is that **scale makes thorough testing of big data applications before deployment** both **impractical and infeasible**.

Even if you could test at deployment scale, as soon **as your data grows,** your **code is operating beyond its tested tolerances**.

The only way to discover if your new components operate correctly is to deploy them on the production system and use techniques such as canary testing to **validate their behavior**.

http://whatis.techtarget.com/definition/canary-canary-testing

The Challenges of Scale-4

Finally**, it's important** to see these challenges of **failure handling**, **complexity**, and **testing** in the context of contemporary **big data system deployments.**

The **scale** of companies like **Netflix**, which uses Cassandra to manage data on 750 nodes as part of its **cloud-based software infrastructure**, are well documented **examples that herald the future** for many large **government and business organizations.**

http://www.datastax.com/resources/casestudies/netflix

Four Principles of Engineering Scalable, Big Data Software Systems

First Principle: System Costs Must Grow More Slowly Than System Capacity

Second Principle: The More Complex a Solution, the Less Likely it Will Scale

Third Principle: Avoid Managing Conversational State Outside the Data Tier

Fourth Principle: You Can't Manage What You Don't Monitor

Final Thoughts and Looking Ahead

The four principles hold for any big data system, **so adhering to them will always be a good thing.**

In contrast, unconsciously **violating these principles** is likely to lead a system into **downstream distress**, **slowing capability delivery**, massively **inflating costs** and potentially **leading to project failure**.

Of course, **simple expediency may mean you have to violate a principle** to **meet a near-term deliverable**. **Violations are not a bad thing**, as long you **recognize the technical debt** that has been **incurred and plan accordingly to pay this debt back before the interest incurred becomes a major problem.**

http://blog.sei.cmu.edu/archives.cfm/category/technical-debt

Addressing the Software Engineering Challenges of Big Data

http://blog.sei.cmu.edu/post.cfm/challenges-big-data-294

A NON-GEEK'S BiG DATA PLAY BOOK

http://www.sas.com/content/dam/SAS/en_us/doc/whitepaper1/non-geeks-big-data-playbook-106947.pdf

Addressing the Software Engineering Challenges of Big Data-2

New data sources, ranging **from diverse business transactions to social media, high-resolution sensors**, are creating **a digital tidal wave of big data** that must be **captured, processed, integrated, analyzed, and archived.**

Big data systems storing and analyzing petabytes of data are becoming **increasingly common** in many application areas. Representing major, long-term investments.

Analysts estimate data storage growth at 30 to 60 percent per year. **Organizations must develop a long-term strategy** to address the **challenge of managing projects** that **analyze exponentially growing data sets with predictable, linear costs.**

This blog post describes a **lightweight risk reduction approach** called **Lightweight Evaluation and Architecture Prototyping** (for Big Data) we **developed** at the SEI.

. http://blog.sei.cmu.edu/post.cfm/challenges-big-data-294

Challenges of Big Data-3

For the **DoD, Military operations, intelligence analysis, logistics, and health care** all represent **big data applications** with **data growing at exponential rates.**

In 2012, the **DoD announced a $250 million annual R&D investment in Big Data** targeted at **specific mission needs** such as **autonomous systems and decision support.**

For example,the **DoD has developed a roadmap through 2018** that identifies the **need for distributed, multi-petabyte data stores** that can **underpin mission needs for scalable knowledge discovery, analytics, and distributed computations**.

The following examples illustrate two different DoD data-intensive missions:

Electronic health records. Providing **care for more than 9.7 million active-duty personnel,** their **dependents, and retirees**. The **15-year-old repository** operates a **continously growing petascale database**, with **more than 100 application interfaces**. System workload types.

Flight data management. **Modern military avionics systems** capture **tens of gigabytes (GBs) of data per hour of operation**. This data is **collected in in-flight data analysis systems**, which **perform data filtering and organization**, **correlation with other data sources**, and identification of significant events. **These capabilities support user-driven analytics**

To address these big data challenges,

A **new generation of scalable data management technologies has emerged in the last five years.**

Relational database management systems providing strong data-consistency guarantees **based on vertical scaling of compute and storage hardware, are being replaced by NoSQL**

These NoSQL databases **achieve high scalability and performance using simpler data models, clusters of low-cost hardware**, and **mechanisms for relaxed data consistency** that enhance performance and availability

To address these big data challenges-2

A **challenging technology adoption problem** is being **created**, by a **complex and rapidly evolving landscape of non-standardized technologies** built on **radically different data models**.

Selecting **a database technology that can best support** mission needs for **timely development**, **cost-effective delivery, and future growth is non-trivial.**

Using these new technologies to **design and construct a massively scalable big data system** creates an **immense software architecture challenge for software architects and DoD program managers.**

Why Scale Matters in Big Data Management

Scale has many **implications for software architecture.**

The first revolves around **the fundamental changes** that scale enforces on **how we design software systems**.

The second is **based upon economics**, where **small optimizations in resource usage at very large scales can lead to huge cost reductions** in absolute terms. These two issues:

Designing for scale.

Economics at scale.

Big Data and Process Improvement

Linking Big Data to Big Process Improvement…An Imperative

http://www.capgemini.com/blog/bpo-thought-process/2014/03/linking-big-data-to-big-process-improvementan-imperative

Model&DeployProcessImprovementsQuickly&SeamlesslywithBPM Tools

http://www.opentext.com/campaigns/business-process-management-bpm-case-management-whitepaper?OM=SEM_BPM_Whitepaper&gclid=Cj0KEQjwvqWgBRChnMjQ7u7UzOUBEiQAooXvYVRGBn4UO7GcO1EU10zLx8g1xyyTJQj2vwk0k-nQIzgaAq4A8P8HAQ

Operationalizing the Buzz:

Big Data 2013

http://www.pentaho.com/sites/default/files/uploads/resources/ema_bigdata_2013_operationalizing_the_buzz_1.pdf

Linking Big Data to Big Process Improvement…An Imperative

http://www.capgemini.com/blog/bpo-thought-process/2014/03/linking-big-data-to-big-process-improvementan-imperative

A recent **Capgemini survey** reveals that *"**organizations are** increasingly **prioritizing big data as an asset**, and once realized the **bottom line improvements can be very noticeable**."* In addition, the survey notes that **decisions are increasingly based on hard analytics**, but **with the massive amounts of data available**, **accessibility** - or the time it takes to produce meaningful analytics - **can substantially slow down the decision-making process** as well as the business, **ultimately** producing a **negative impact on the bottom line**.

To make sense of the relationships between the data and an organization's business processes, it **takes creativity, statistical prowess, continuous improvement expertise, and process knowledge**, the same factors necessary to validate a root cause. **Not using any one of these elements means the enterprise-wide potential of Big Data and Big Process improvement is delayed, or anemic in the production of real results.**

Model&Deploy Process Improvements Quickly&Seamlessly with BPMTools

http://www.opentext.com/campaigns/business-process-management-bpm-case-management-whitepaper?OM=SEM_BPM_Whitepaper&gclid=Cj0KEQjwvqWgBRChnMjQ7u7UzOUBEiQAooXvYVRGBn4UO7GcO1EU10zLx8g1xyyTJQj2vwk0k-nQIzgaAg4A8P8HAQ

The **success of your company's operations** is **contingent on strategically aligning the business with IT.**

Download OpenText's Business Process Management Whitepaper to see how you can accomplish that & improve your ROI.

Operationalizing the Buzz:

Big Data 2013

http://www.pentaho.com/sites/default/files/uploads/resources/ema_bigdata_2013_operationalizing_the_buzz_1.pdf

Operationalizing the Buzz: Big Data 2013

Model&DeployProcessImprovementsQuickly&SeamlesslywithBPM Tools

The **success of your company's operations** is **contingent** on **strategically aligning the business with IT.**

Download OpenText's Business Process Management Whitepaper to see how you can accomplish that & improve your ROI.

Operationalizing the Buzz: Big Data 2013

An ENTERPRISE MANAGEMENT ASSOCIATES® (EMA™) and 9sight Consulting Research Report

November 2013

EMA

IT & DATA MANAGEMENT RESEARCH,

INDUSTRY ANALYSIS & CONSULTING

Machine Learning

http://en.wikipedia.org/wiki/Machine_learning#Definition

Definition

In 1959, Arthur Samuel defined machine learning as a "**Field of study** that gives **computers** the **ability** to learn **without being explicitly programmed".**

Tom M. Mitchell provided a **widely quoted**, **more formal definition**: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E". This **definition** is notable for its **defining machine learning in fundamentally operational rather than cognitive terms**, thus following Alan Turing's proposal **in Turing's paper "Computing Machinery and Intelligence"** that the **question** "**Can machines think**?" be replaced with the question "**Can machines do what we (as thinking entities) can do?**"

Machine Learning 2

Generalization

A core objective of a learner is to generalize from its experience.[
Generalization in this context is the **ability of a learning machine to perform accurately** on new, **unseen examples/tasks after having experienced a learning data set**.

The **training examples** come from **some generally unknown probability distribution** (considered representative of the space of occurrences) and the **learner has to build a general model about this space** that **enables** it to **produce sufficiently accurate predictions in new cases.**

Machine Learning 3

Machine learning and data mining

These **two terms are commonly confused**, as they often employ the same methods and overlap significantly. They can be roughly defined as follows:

Machine learning focuses on prediction, based on ***known*** **properties learned from the training data.**

Data mining focuses on the discovery of (previously) ***unknown*** **properties in the data.** This is **the analysis step of Knowledge Discovery in Databases.**

Human interaction

Some **machine learning systems attempt to eliminate** the need for **human intuition** in data analysis, while **others adopt a collaborative** approach between human and machine.

Machine Learning at CMU

http://www.ml.cmu.edu/

What is the Machine Learning Department?

The Machine Learning Department is an academic department within Carnegie Mellon University's School of Computer Science. We focus on research and education in all areas of statistical machine learning. Watch an interview with Tom Mitchell, Department Head:

Interview with Tom Mitchell

http://videolectures.net/mlas06_mitchell_itm/

CMU Initiatives involving Big Data

The Simon Initiative

http://www.cmu.edu/leadership/president-suresh/presidential-communications/2013-11-11.html

Brain, Mind & Learning

http://www.cmu.edu/research/brain/people/

Brain Hub

http://www.cmu.edu/research/brain/

Global Learning Council

http://www.cmu.edu/news/

Events

http://globallearningcouncil.org/events/

http://www.cmu.edu/simon/docs/GlobalLearningCouncil.pdf

The Simon Initiative

http://www.cmu.edu/leadership/president-suresh/presidential-communications/2013-11-11.html

Brain, Mind & Learning

http://www.cmu.edu/research/brain/people/

People

The individuals representing a snapshot of the multitude of Carnegie Mellon researchers who are pioneers in the field of brain, mind and learning:

Brain Hub

http://www.cmu.edu/research/brain/

BrainHub[SM]: Harnessing the technology that helps the world explore brain and behavior

Global Learning Council

http://www.cmu.edu/simon/docs/GlobalLearningCouncil.pdf

The world is in the midst of a revolution in education. Connected through technology, learners everywhere may have free access to content online, creating new learning pathways both inside and outside traditional educational institutions.

Events

http://globallearningcouncil.org/events/
http://www.cmu.edu/simon/docs/GlobalLearningCouncil.pdf


Contact Information

Angel Jordan

Carnegie Mellon University

412-268-2590

ajordan@cs.cmu.edu

U.S. Mail

Institute for Software Research

Carnegie Mellon University, Wean Hall 5309

5000 Forbes Avenue

Pittsburgh, PA 15213

USA

This material is based upon work funded and supported by the Department of Defense under Contract No. FA8721-05-C-0003 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.