Preliminary Report for the PDS Architecture 2008+

Architecture WG
(Acton, Crichton, LaVoie, Martin, Stein)

December 4, 2007

Outline

- Introduction
- Core Concepts and Background
- Drivers
- Core Architectural Principles
- PDS Architecture Concept
- Decomposition of the elements
- Existing Gaps in the requirements
- Answers to PDS4 Questions
- Management Council Recommendations

Introduction

- PDS4 Architecture WG focus on PDS overall System Architecture
  - Core Processes
  - Data Architecture
  - Technology Architecture

- WG followed the following process
  - Evaluation of the PDS Roadmap
  - Evaluation of the PDS Level 1,2,3 Requirements
  - Construction of of a set of architectural drivers
  - Identification of the elements of the PDS4 System Architecture
  - Final report which includes recommendations to the MC on an initial implementation plan

Core Concepts and Background

- Architecture: The fundamental organization of a system embodied in its components, their relationships to each other, and to the environment, and the principles guiding its design and evolution. (ANSI/IEEE Std. 1471-2000)

- PDS4 Reference System Architecture is decomposed into three core pieces:
  - Process Architecture
    - Describes the core processes PDS follows for its system
    - PDS examples: archive management, preservation planning, peer review, standards management, etc
  - Data Architecture
    - Describes the information models and data standards PDS follows for its system
    - PDS examples: PDS data model, PDS data dictionary, ODL (Grammar), etc
  - Technology Architecture
    - Data management, storage, tools, portals, etc

- The WG used this to understand how to decompose the system and then plan for its evolution

Drivers

- The PDS Roadmap provides a number of drivers for the next ten years
  - The WG extracted these drivers and identified a set of related architectural drivers
- The PDS Management Council provided a number of questions to be answered by PDS4 Working Groups
  - The WG developed a set of responses to each of the questions which we will tie to our PDS4 architecture concept

Summary of PDS Architectural Drivers

- More Data: PDS storage requirements are projected to increase from 40 TB to over 500 TB in just three years. This will require more automation, scalable high capacity storage systems and advanced data movement techniques.

- More Complexity: Missions, instruments, and data are all becoming more complex. This will require an improved information model for archiving diverse data products (in situ, geographical, astronomical) as well as a modern online data dictionary with name space management and access control.

- More Producer Interfaces: PDS is facing an increasing number of missions, a greater number and diversity of data providers, and smaller, focused missions. This will require a streamlined standards architecture that is easy to learn and use, with more reliance on delivering data in standard data formats. Cross-platform archiving tools must be provided which can be used to design, generate, validate, and deliver archival data sets.

Summary of PDS Architectural Drivers

- Greater User Expectations: The World Wide Web has led users to expect well-documented data to be readily available via text-based or graphical search systems with data delivery in a variety of formats compatible with their data processing systems. This includes access to tools for displaying or analyzing discipline specific data as well as special processing to produce higher order products.

- Limited Funding: The emphasis on smaller, faster, cheaper missions which often include international partners may limit the ability to provide products suitable for analysis by the broader science community. This puts a burden on NASA Data Analysis programs or on the PDS have to finish the job. As space exploration continues to become an international effort, PDS must expend increasing resources working with foreign agencies and international organizations to assure access to new mission data. The "internationalization" of space exploration will also necessitate additional standards that promote data sharing and interoperability and an international core data model for archiving and for querying remote archives.

Summary of PDS Architectural Drivers

- Creating a "system" from the federation: The current PDS nodes operate autonomously and independently with limited distributed access via PDS-D to node repositories. This means that each site must do its own planning, design, review, procurement, code development, testing and operations. There is little sharing of technical expertise in this heterogeneous environment. A better approach would be to provide technology specifications to allow distributed and shared services across the federation, and to ensure that tools can plug into local environments. Common infrastructure services would be provided where it makes sense (physical media production, security, backup, mirroring, web site maintenance).

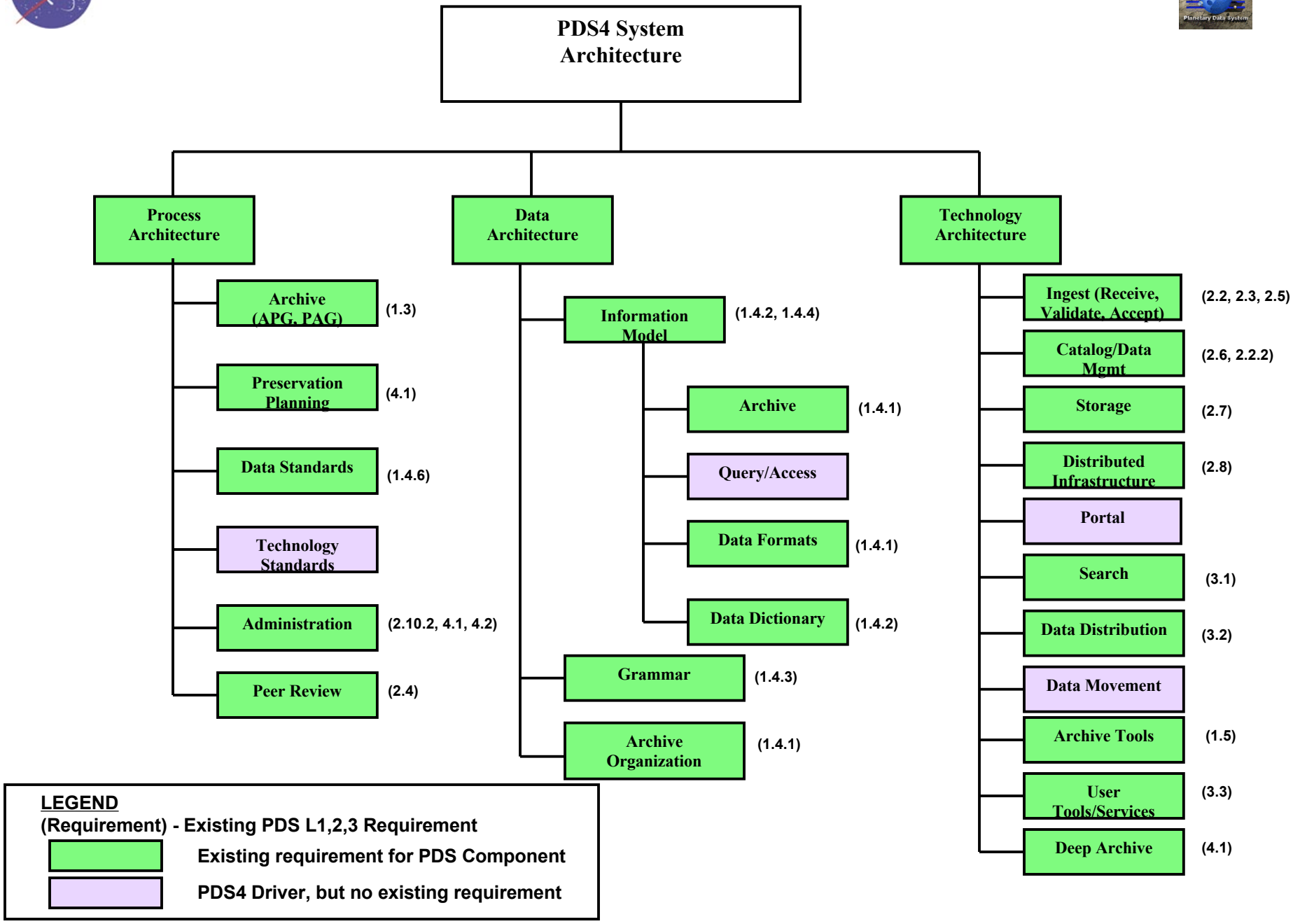
Core Architectural Principles

- Model driven
  - The system is based on the model
- Archiving is the priority
  - The system is designed with archiving as the priority
- Evolution of the system as elements
  - The system has a modular architecture allowing for independent evolution of elements
- Support for a distributed federation
  - Highly distributed allowing changes in federation structure and rules
- Use of standards
  - Standards are rigorously used. PDS adopts before developing, where possible
- Low cost of ownership
  - PDS ensures data providers and nodes can adopt and use tools with minimal resource impact
- Diversity
  - PDS is designed to suport diverse needs of providers, missions and planetary science community
- Scalability
  - PDS is designed to scale core functions of the system
- Explicit Design
  - Elements of the system are explicitly defined with unambiguous specifications
- International Adoption
  - Standards and tools are defined and implemented in order to allow for international adoption
- Integrity
  - Data integrity is architected into PDS processes and the system end-to-end
- Timeliness
  - PDS works with data providers as early as possible to adopt processes, standards and tools

PDS4 Architecture Concept

- PDS4 is explicitly architected as an online distributed system
- As an online system, PDS4 shall
  - Fully implement the recommendations of the PMWG with primary, secondary and archive copies of data
  - Deliver all data to the NSSDC
- The PDS4 service architecture shall
  - Enable server-side processing and access to data and catalogs
  - Fully support PDS 2.8 requirements
  - Be based on standards (protocols, interfaces, operating systems, development environments, etc)
  - Ensure the distributed infrastructure supports all higher functions (ingest, search, process, deliver)
  - Provide software and guidelines for developing services

Decomposition of PDS System Architecture



Gaps in our requirements

- PDS4 identified architecture elements with no level 3 requirement
  - Technology Standards
  - Query/Access Models
  - Portal
  - Data Movement (packaging and transport)

- PDS4 questions which “suggest” a need for a level 3 requirement
  - One-stop shopping/seamless access to data
  - Support discipline-specific and product specific server-side processing
  - Providing libraries for core archive functions which enable construction of both PDS and non-PDS tools
  - Data integrity support for data transfer and delivery
  - Standard software and guidelines for building online distributed services

- Recommend that PDS update Level 1/2/3 requirements and baseline for PDS4

AWG Responses to PDS4 Questions

QUESTION	PDS4 ARCHITECTURE WG RESPONSE	SOURCE
How will PDS-4 enable "one-stop shopping", I.e., seamless access to data that reside at multiple nodes? How will PDS- help the user to "locate" data of interest (or accurately conclude they are not available in the system)?	PDS4 architecture shall support delivery of products to the user from distributed repositories without the user's knowledge of the data location. Additionally, the PDS4 architecture shall allow for discovery of data (to the product level) across nodes, however, it requires that standards be applied and that the system provide consistent results of holdings across PDS based on an integrated search architecture. We are not implying that the distributed federation disappears or that we have a single interface to all data, but rather we better integrate it. The architecture team proposes that a level 3 requirement be added and a project be started to address the search architecture which includes the process, data, and technology specifications to enable this.	Arvidson / Simpson
How will PDS-4 help users by delivering derived data products in the format, coordinate system, and map projection from the user requests? How will PDS-4 help users to **create** derived data products from raw and/or calibrated archives? Since most data are delivered in raw form, what are we doing to improve the user's ability to perform calibration and other processing before reaching the display stage?	There are three cases in which derived data products and data sets are created/modified, in order of preference: •By the data provider (provider side) •Prior to delivery to the user (server side); shall provide standard services that run at the nodes that can operate on the data to deliver them in the form required •By the user after data delivery(client side); this may include user-provided tools, PDS-provided standalone tools, PDS-provided API All three cases do not necessarily apply to every data set.	Arvidson / Simpson

AWG Responses to PDS4 Questions(2)

How will PDS-4 help data providers by automating the design, production, and delivery of PDS data sets?	Automation of PDS is a key recommendation by the PDS4 Architecture WG and one that is critical for responding to the architectural drivers.	Arvidson
How will PDS-4 ensure that PDS standards are simple, straightforward, and consistent so that data providers and users can easily understand [**and uniformly**] apply them?	The PDS4 Architecture WG recommends that the standards be explicit, verifiable, and consistent so that they can be implemented both domestically and internationally.	Arvidson (w/ Simpson Modification)
Should we default to machine validation of everything except science content? Then our standards can be very brief; the real test is whether data products pass the validation. What are the risks in terms of loophole discovery and exploitation?	A core principle of the architecture is “model driven”. In other words, the tools should be able to verify that PDS data products are consistent with the model. However, as mentioned, it means the model needs to be explicitly defined using rigorous computer science modeling techniques.	Simpson

AWG Responses to PDS4 Questions (3)

How will PDS-4 ensure that data sets can be safely and efficiently archived in NSSDC and retrieved on demand?	The PDS4 Architecture WG believes that electronic delivery to to NSSDC is a critical piece of having an online system. It recommends that PDS4 implement the recommendation of the PMWG from the August 2007 meeting where there are three repositories for PDS data (primary, secondary and NSSDC).	Arvidson
How will PDS-4 improve the data transfer, data integrity, and maintenance of PDS data sets?	The PDS4 Architecture WG recommends that the data integrity requirements be adopted and that an implementation plan be developed for them. In addition, PDS should define technical standards and solutions for the movement of data across the PDS enterprise (packaging and transfer).	Arvidson
How will PDS-4 Improve the monitoring of data ingestion which takes place over an extended time? Is “new-CATS” integral to PDS-4?	End-to-end tracking from data providers to the deep archive is currently defined as part of the PDS data integrity policy and is included in the draft data integrity requirements. As mentioned above, the PDS4 Architecture WG recommends that implementation of data integrity be a priority project for PDS.	Simpson
How will PDS-4 improve the automated management of the archive so that, once ingested the data are easily relocated and retrieved without requiring human intervention? What are we doing to ensure that our computers do MORE of the routine work?	One of the benefits of coupling an online system with a distributed service infrastructure, is allowing data products to move around the PDS network and still be located, accessed and distributed. “Location independence” should be a requirement for PDS.	Simpson

AWG Responses to PDS4 Questions (4)

What about PDS-4 will simplify addition of future user services -- for example, the hypothetical "geometry engine"? Do we have robust building blocks at the foundation of our structure to that it is easy to grow services that we haven't yet imaged?	The PDS4 Architecture concept is a distributed, service-oriented architecture which provides "hooks" for plugging in services across the enterprise. These are services that can be collocated with appropriate computing support. The PDS4 Architecture WG recommends that PDS provide a standard set of software and guidelines for building online data services.	Arvidson
How will PDS-4 improve our ability to document and/or correct errors in data sets which have completed the ingestion process … or to add to data set metadata (the dynamic master index)?	The PDS4 Architecture WG largely believes that update practices should be defined in a policy. It does, however, recognize that there needs to be an appropriate process and mechanism in place to support updates along with the addition of new metadata.	Simpson
Should PDS-4 be required to be backwards compatible?	KEY QUESTION It may not be possible to be fully backward compatible, however, efforts should be made to ensure that all of PDS data can be located and used regardless of the original version in which is was captured. Additionally, PDS4 should be designed to be forward compatible.	Sykes

AWG Responses to PDS4 Questions (5)

What are the costs if it is or is not (in terms of maintaining two archives, retrofitting old data sets, IPDA issues, …)?	The PDS4 Architecture WG does not believe you can build another data system and run them in parallel without an appropriate wedge in the budget to support it. The WG recommends a phased approach by identifying how each architectural element can be moved forward to PDS4.	Simpson
How will PDS-4 enable users to find the specific data they need - down to the product level?	See question 1.	Gordon

Summary of MC Recommendations (from PDS4 Questions)

- Update PDS level 3 requirements addressing gaps
- Initiate project on defining the overall search architecture to enable one-stop shopping and seamless access for PDS
- Adopt PMWG recommendations as an all online system with three electronic copies of the data
- Define technical standards and solutions for the movement (packaging, transport) of data across the PDS enterprise
- Adopt the data integrity requirements and initiate a project for its implementation
- Provide a core set of software and guidelines for building online data services

Backup

PDS4 Architecture Concept Cont…

- Data Integrity
  - Fully implement data integrity across the system through to the NSSDC
  - Develop checksum standards to safeguard against file corruption
  - Implement tracking of data from data providers through to NSSDC
  - Ensure safeguards are in place so data can be accessed and is not lost
- Data Movement
  - Provide standards for the movement of data across the PDS (both network and offline)
  - Adopt and implement critical services to enable high capacity exchange of data between nodes (discipline nodes, data nodes, etc), NSSDC, and secondary repository
  - Currently no requirements or standards exist within PDS
- Archiving Tools
  - Fully implement the archiving tool requirements identified in 1.5.x
  - Develop next generation tool for “display” (3.3.2)?
  - Provide modular tools which can be plugged into node-specific environments
- Automation
  - The system shall provide “hooks” to support automation of critical PDS elements such as ingestion, data integrity checking, etc

PDS4 Architecture Concept Cont…

- Search
  - Develop an explicit search architecture for searching across the entire PDS archive based on the distributed service architecture
    - Develop corresponding requirements for search
  - Allow for specialization of the search architecture at the nodes based on DN search data models
  - Address incompatibilities in products and metadata to enable comprehensive search
- Portals
  - Provide an integrated “portal” architecture that integrates with the search architecture and enables content management, news management, and other Web 2.0 capabilities
  - Support deployment of portal architecture across PDS along with associated standards for sharing content

PDS4 Architecture Concept Cont…

- User Tools and Services
  - Ensure the architecture allows for construction of tools that can be “plugged” into PDS
  - Allow non-PDS tools to interact with the PDS distributed infrastructure
  - Enable access via client APIs that support core industry standards for various disciplines
  - Promote a standard set of core PDS analysis tools for working with PDS data

- Standards
  - Identify BOTH data and technology standards
  - Technology standards should guide system interfaces
  - Data standards should guide data definitions
  - Ensure standards are straightforward, explicit and unambiguous allowing for use in implementations by both PDS and non-PDS personnel

Final Conclusions

- The WG believes that it's important to think about the elements of the system and their evolution
  - It allows us to phase movement towards a PDS4 target
  - Reduces risk
  - Some elements may require little change for PDS4

- The WG believe PDS4 should have an explicit architecture
  - Explicit system architecture with interface specifications
  - Explicit data architecture with data models captured using modern modeling tools
  - PDS needs standards for both data and technology elements
  - Existing processes should be re-examined, however, they should be applicable to PDS4
    - Examples in PDS process documents will need to change to be consistent with changes in PDS standards (data, technology)