IBM Life Sciences

IBM Life Sciences Framework

Developing an integrated, flexible infrastructure for life sciences research and development

ITSO iSeries Technical Forum
RP01

Marcela Adan

IBM Life Science Framework Development

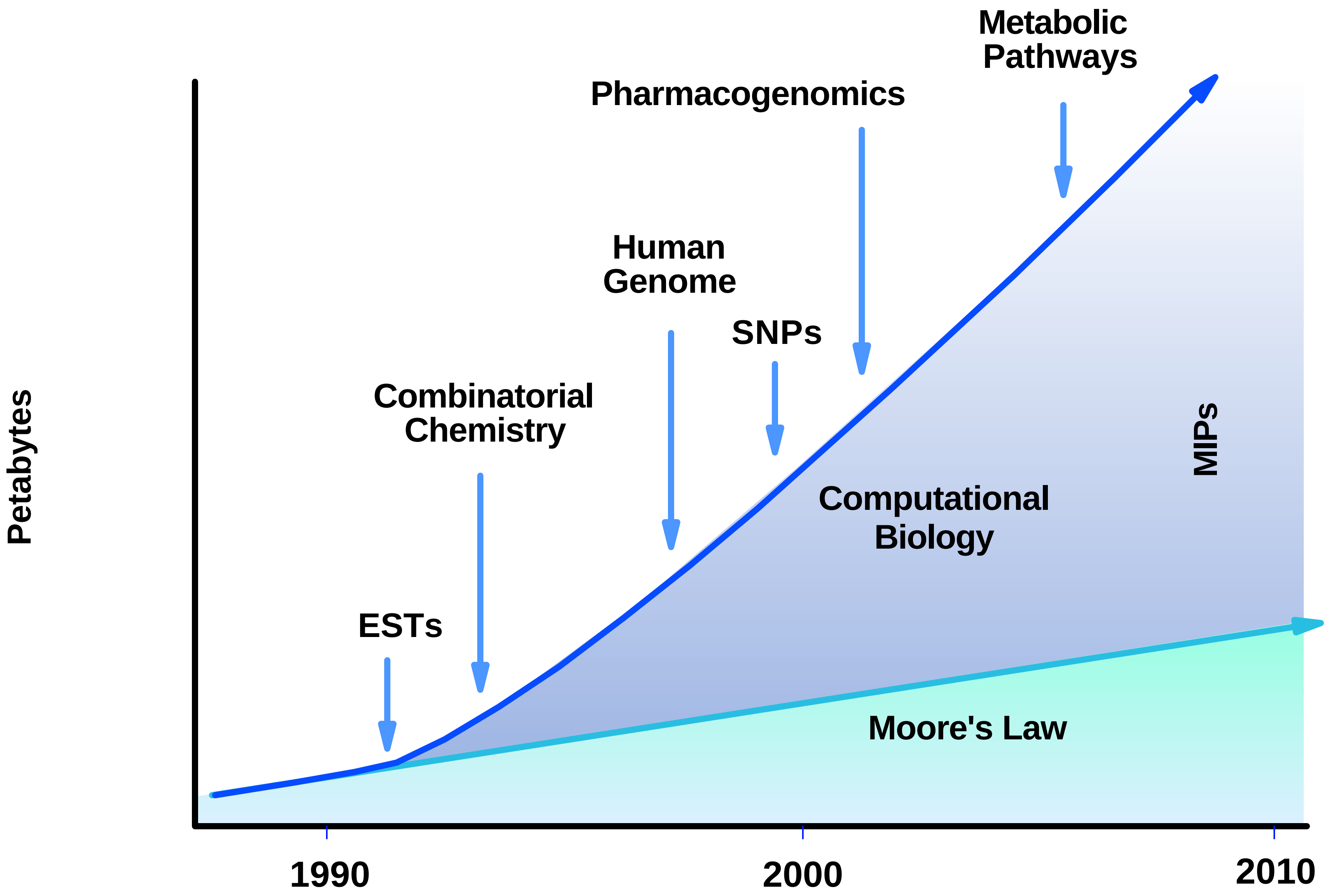
Agenda

- The Life Sciences challenges
- Life Sciences Technologies from IBM
- LS Framework Overview
  - Open Solution to the LS problems
  - Framework architecture
  - Framework Technologies
- Proof of Concept, Pilots & Future Work

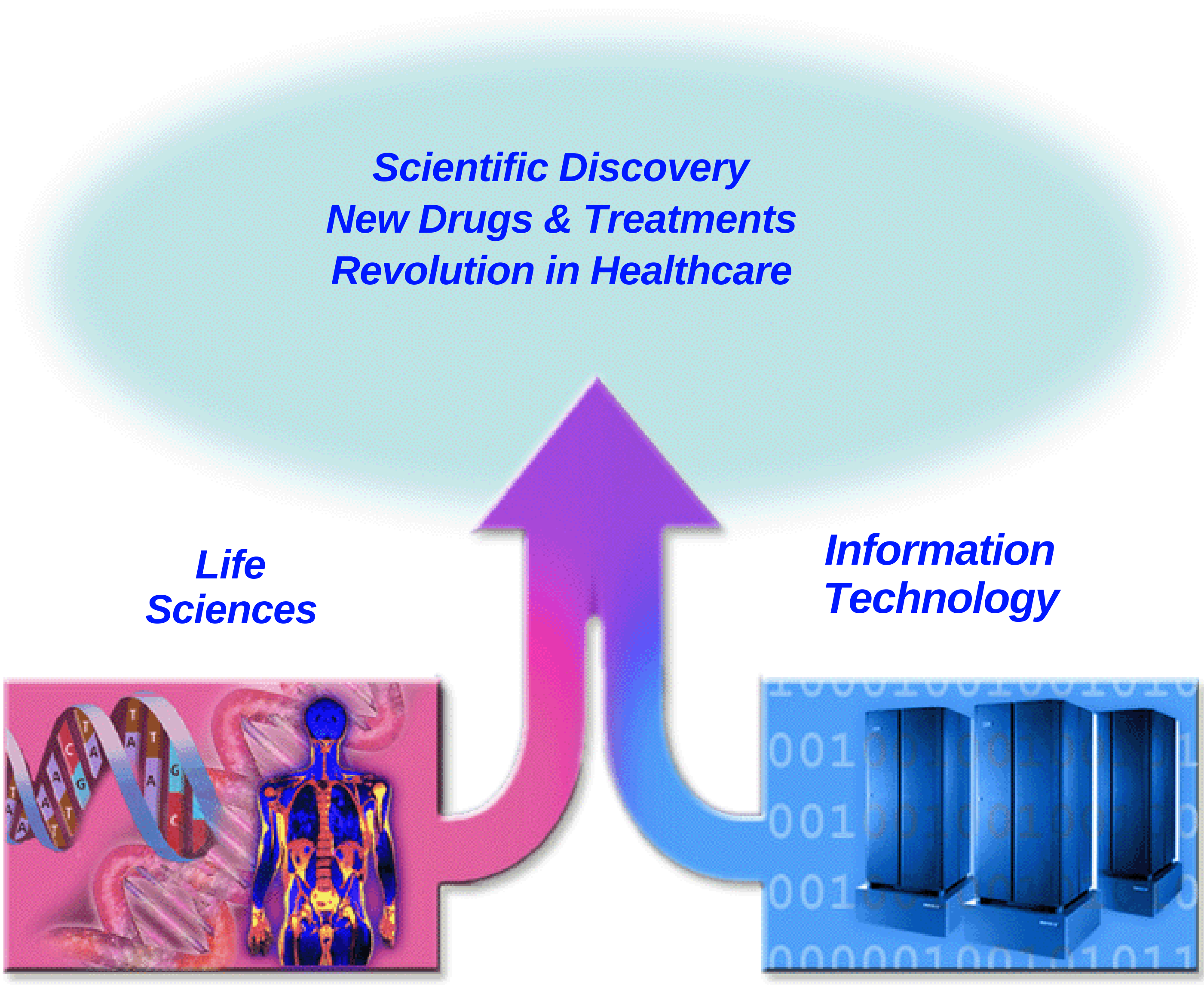
Life Sciences Information Technology focus areas

- **Genomics/Proteomics**
  - Harnessing the true power of data through integration, visualization, and prediction (better information, more tools, improved leverage)
- **Drug Discovery / Cheminformatics**
  - Streamlining the discovery process by eliminating bottlenecks and embracing collaboration
- **Clinical Trials**
  - Reduced cycle times, improved data management for cost efficiencies, increased numbers of products to market
- **Medical Informatics**
  - Help institutions to develop the most effective drugs and treatments based on an individual's genetic and phenotypic characteristics using information-based medicine

Life Sciences data management requirements are growing faster than Moore's Law



The success of life sciences R & D depends on the convergence of IT and science



Scientific Discovery

New Drugs & Treatments

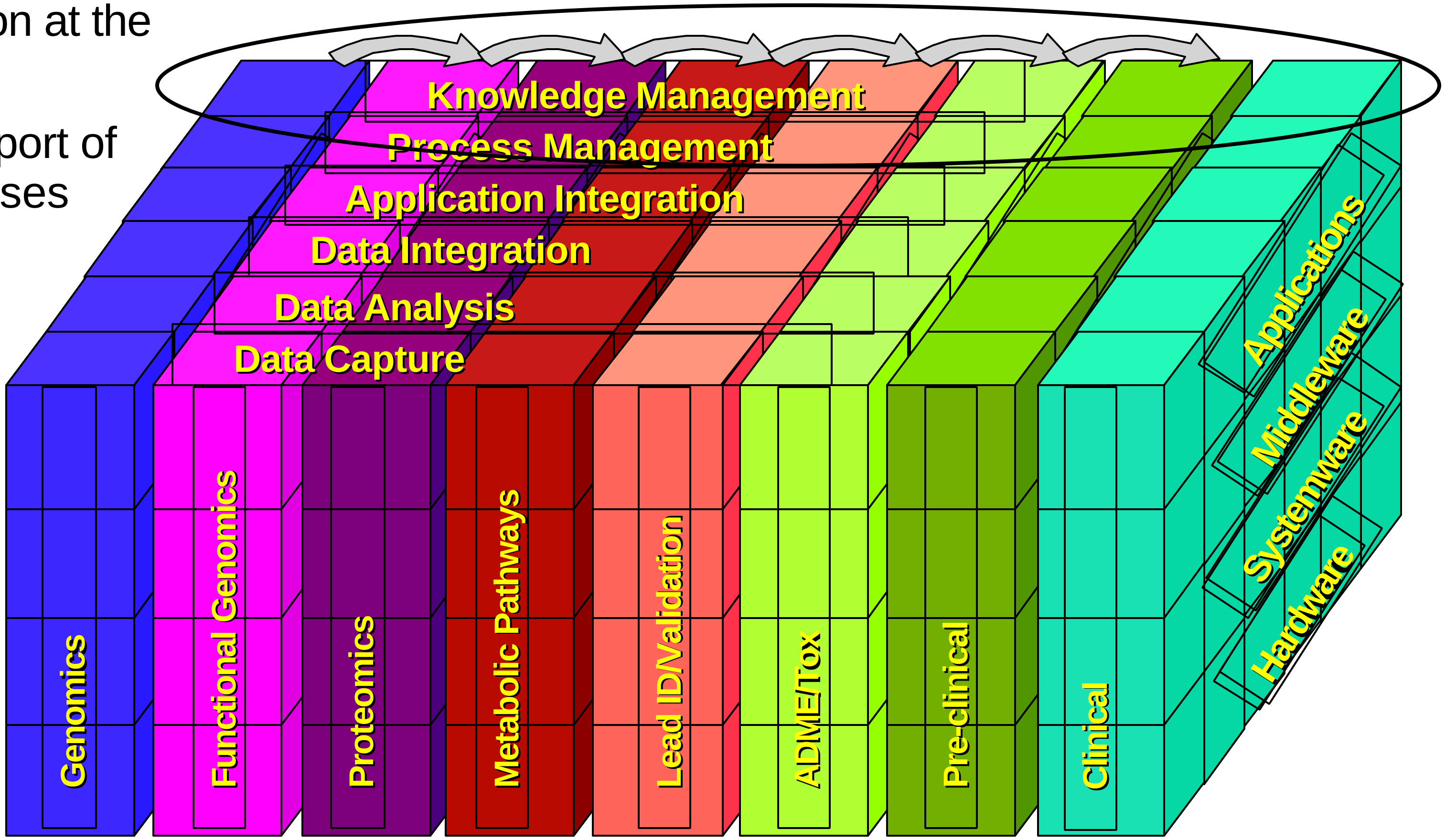
Revolution in Healthcare

Life Sciences

Information Technology

Today's Life Sciences R&D IT System

- Vertically organized
- Restricted sharing
- Manual integration at the user interface
- Inconsistent support of research processes



Challenges facing Life Sciences R&D organizations

- Accessible and secure integration of increasing and diverse data sources, internally and externally
- Integration of applications across different R&D functional areas
- Knowledge management, sharing and collaboration
- Data management, security, access, and storage management
- Business-to-business integration for outsourced functions

IBM Life Sciences Framework

- The environment where IBM and industry providers help customers accelerate the transformation of their life sciences R&D IT systems.
- This environment is built on an infrastructure of industry standards, proven technologies and methodologies, supporting openness to enable the integration of domain-specific functions.
- IBM, in conjuction with leading life sciences providers, uses this infrastructure to deliver the critical solutions required to create a collaborative research centric environment to improve the drug discovery process.

Challenge: Integration of increasing and diverse data sources

Issues:

- Multiple data sources
- Lack of common representation of data
- Different / inconsistent access control and auditability
- Inability to use visualization tools against various applications' data simultaneously

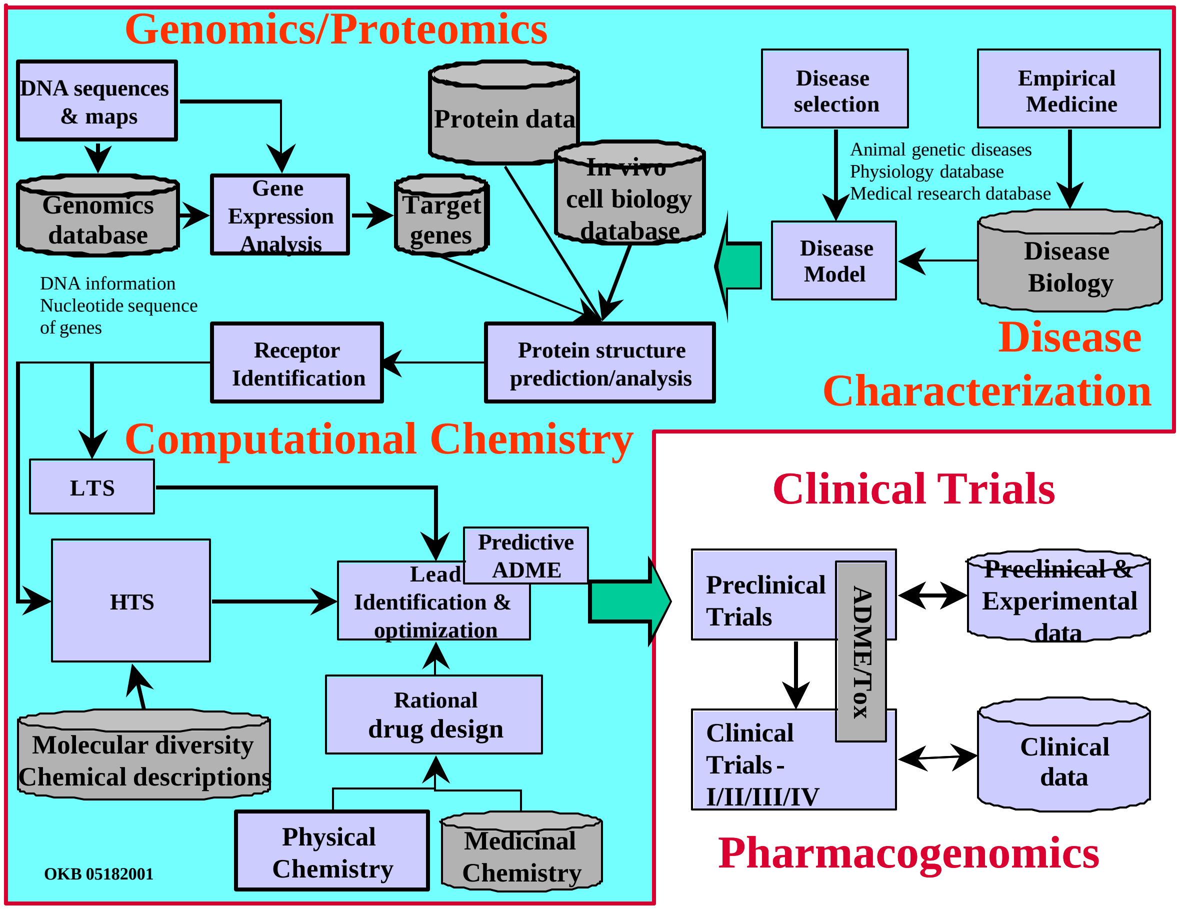
Solution

- Provide a unified view of cross-discipline data using:
  - Data Federation
  - Relational database engines
  - Data source wrappers
  - Data Mining for text
  - Visualization of complex data and its relationships

Benefit

- Provides greater insight with an aggregated view
- Saves time, reduces effort / error
- Leverages critical human resources
- Increases laboratory productivity and efficiency
- Enables collaborative research across companies

Challenge: Integration across different functional areas within the R&D organization



Genomics/Proteomics

DNA sequences & maps

Genomics database

Gene Expression Analysis

Protein data

Target genes

In vivo cell biology database

DNA information
Nucleotide sequence
of genes

Receptor Identification

Protein structure prediction/analysis

Disease selection

Empirical Medicine

Animal genetic diseases
Physiology database
Medical research database

Disease Model

Disease Biology

Disease Characterization

Computational Chemistry

LTS

HTS

Predictive ADME

Lead Identification & optimization

Rational drug design

Molecular diversity
Chemical descriptions

Physical Chemistry

Medicinal Chemistry

OKB 05182001

Clinical Trials

Preclinical Trials

ADME/Tox

Preclinical & Experimental data

Clinical Trials - I/II/III/IV

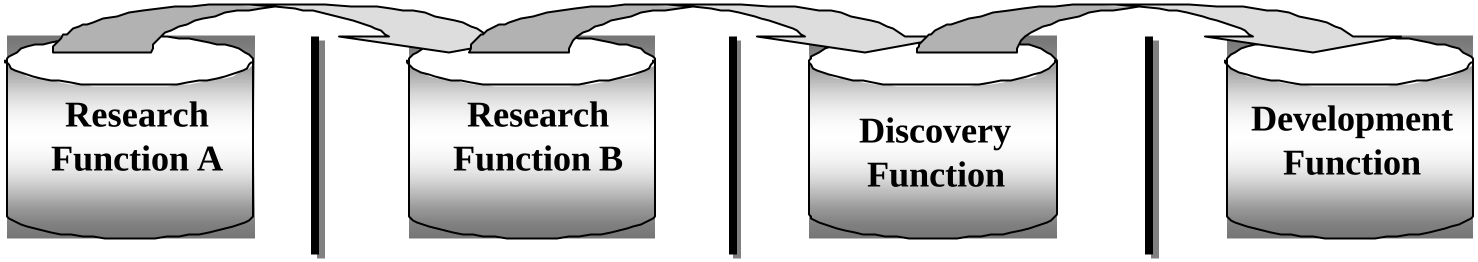
Clinical data

Pharmacogenomics

Issues:

- Manual processing and formatting input and output files for different applications
- Scientists writing code themselves to integrate tasks
- Lack of cross-silo synchronization of applications and data

Challenge: Integration across different functional areas within the R&D organization



Solution

- Publish functional interfaces using web services or other workflow interfaces:
  - Web Application Server (eg, WebSphere)
  - Messaging and Workflow facilities (eg, MQ Series product family)

Benefit

- Unlocks functions trapped in departmental systems
- Minimizes errors due to multiple data entry
- Reduces cost
- Helps shorten time to market
- Transforms processes
- Increased throughput through automated processing

Challenge: Knowledge management, sharing and collaboration

Issues:

- Self-contained organizations impede information sharing
- Overload due to volume of personally non-relevant information
- Cross-organizational insights are not easily accessible
- Organizationally- and geographically-dispersed expertise not fully leveraged

Solution

- Integrated access to customized knowledge, information, and expertise across processes and disciplines
  - Portal Server (eg, WebSphere Portal Server / Lotus K-Station)
  - Knowledge Management Server (eg, Knowledge Discovery Server)
  - Document Management (eg, IBM Content Manager)

Benefit

- Timely access to all information without reformatting or summarization delays
- Enables researchers to act as more cohesive teams

Challenge: Data management, security, access, and storage management

Issues:

- Inconsistent handling and protection of data
- Multiple logons required
- Productivity constraints due to inability to deal with data growth

Solution

- Integrated solutions providing highly available, secure, scalable, and cost effective storage of confidential data
  - Reliable processors and storage
  - Robust operating systems that enable scaling (eg, AIX™, Linux, Solaris)
  - Management tools to monitor and control the environment and security to enforce policies (eg, Tivoli)
  - Application servers to provide domain-specific logic (eg, WebSphere)

Benefit

- Common implementation of data management policies
- Consistently secured and protected data, with cross-discipline access
- Growth unconstrainted by IT limitations

Challenge: Integration for outsourced R&D functions

Issues:

- Systems don't support complex interactions between companies
- User interfaces vary between similiar desktop applications
- Slow, error prone, non-repeatable manual transactions

Solution

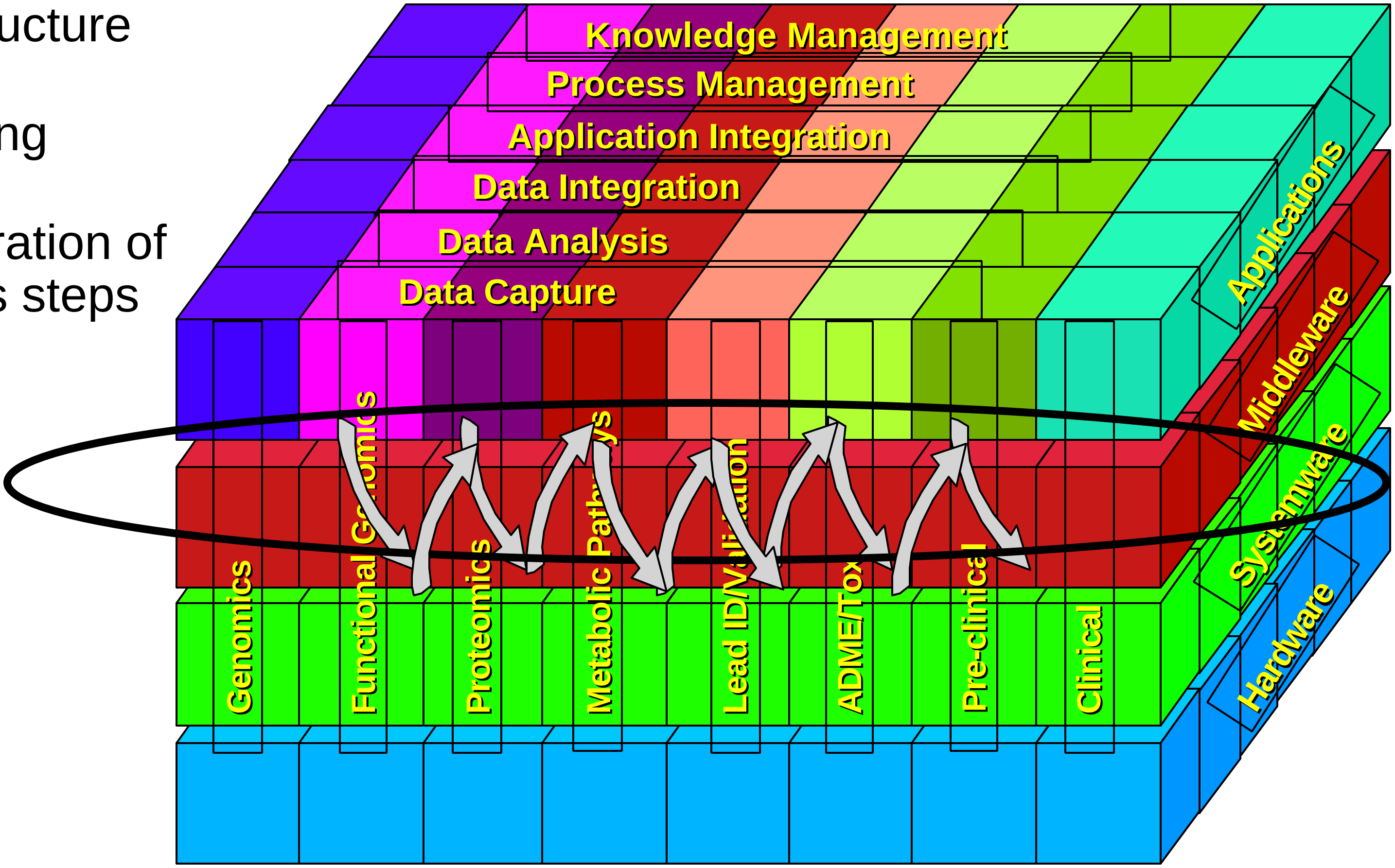
- Leverage web technologies to create common application interfaces using workflow management and guaranteed data delivery
  - Application Server (eg, WebSphere)
  - Workflow (eg, MQ Series)
  - Grid Technology

Benefit

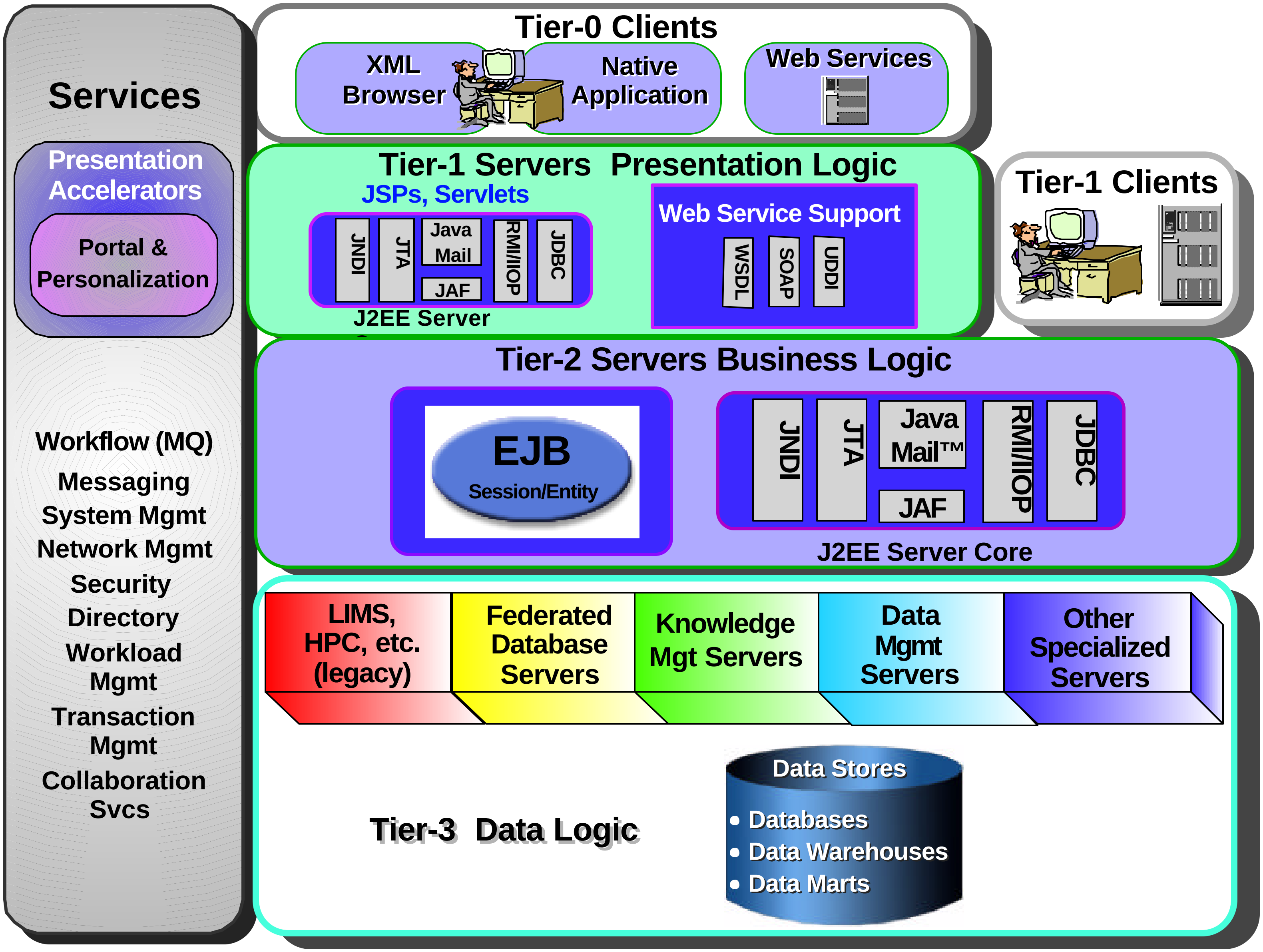
- Immediate access to information or procedures across corporate boundaries
- Reduce chance of errors and delays through accelerated, repeatable steps
- Allow to concentrate on core competencies and better leverage outside expertise

The IBM Life Sciences Framework enables a more collaborative research environment

- Common infrastructure
- Systematic sharing
- Automated integration of research process steps



Life Sciences Framework Architecture



Services

Presentation Accelerators

Portal & Personalization

Workflow (MQ)
Messaging
System Mgmt
Network Mgmt
Security
Directory
Workload Mgmt
Transaction Mgmt
Collaboration Svcs

Tier-0 Clients

XML Browser

Native Application

Web Services

Tier-1 Servers Presentation Logic

JSPs, Servlets

JNDI | JTA | Java Mail | JAF | RMI/IIOP | JDBC

J2EE Server

Web Service Support

WSDL | SOAP | UDDI

Tier-1 Clients

Tier-2 Servers Business Logic

EJB
Session/Entity

JNDI | JTA | Java Mail™ | JAF | RMI/IIOP | JDBC

J2EE Server Core

LIMS, HPC, etc. (legacy)

Federated Database Servers

Knowledge Mgt Servers

Data Mgmt Servers

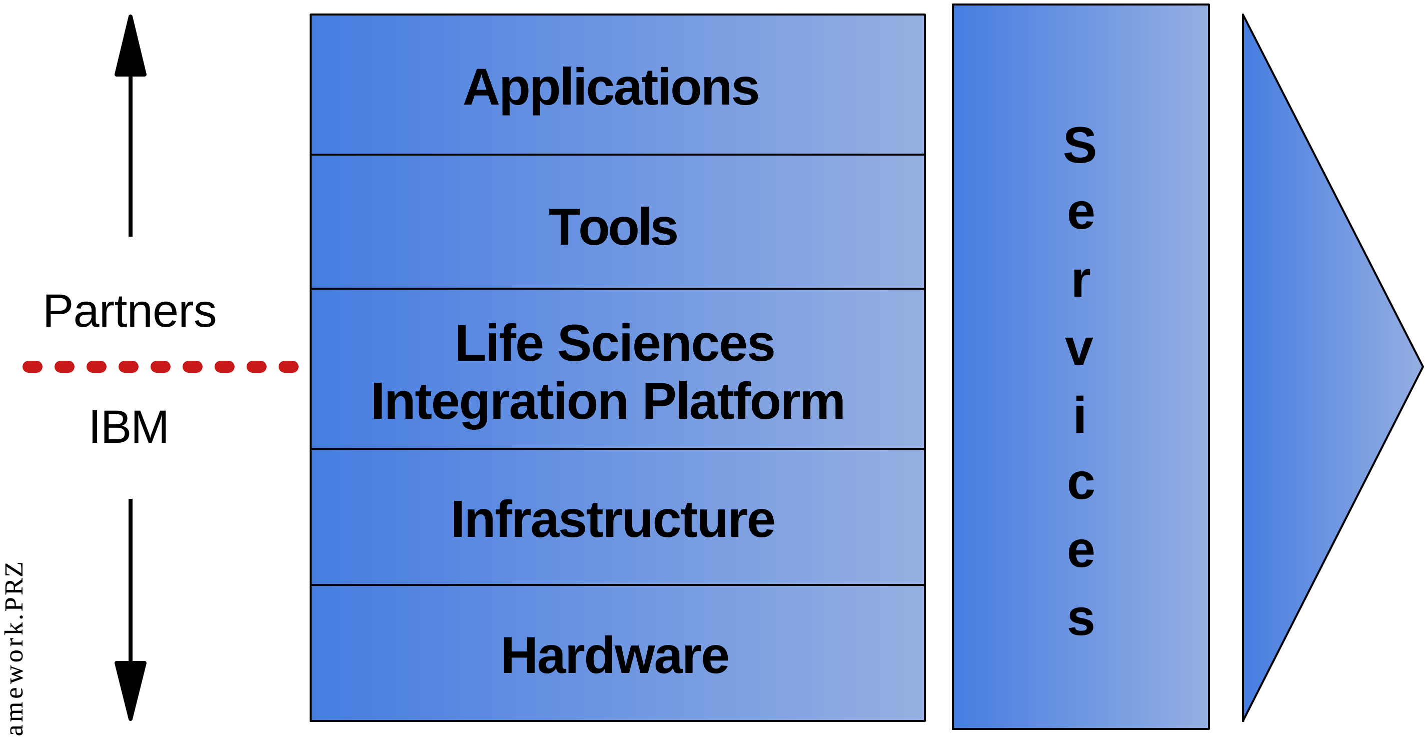
Other Specialized Servers

Tier-3 Data Logic

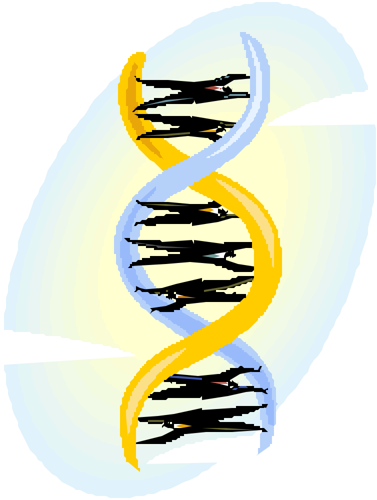
Data Stores

- Databases
- Data Warehouses
- Data Marts

IBM is teaming with leading industry solution providers to create the Life Sciences Framework

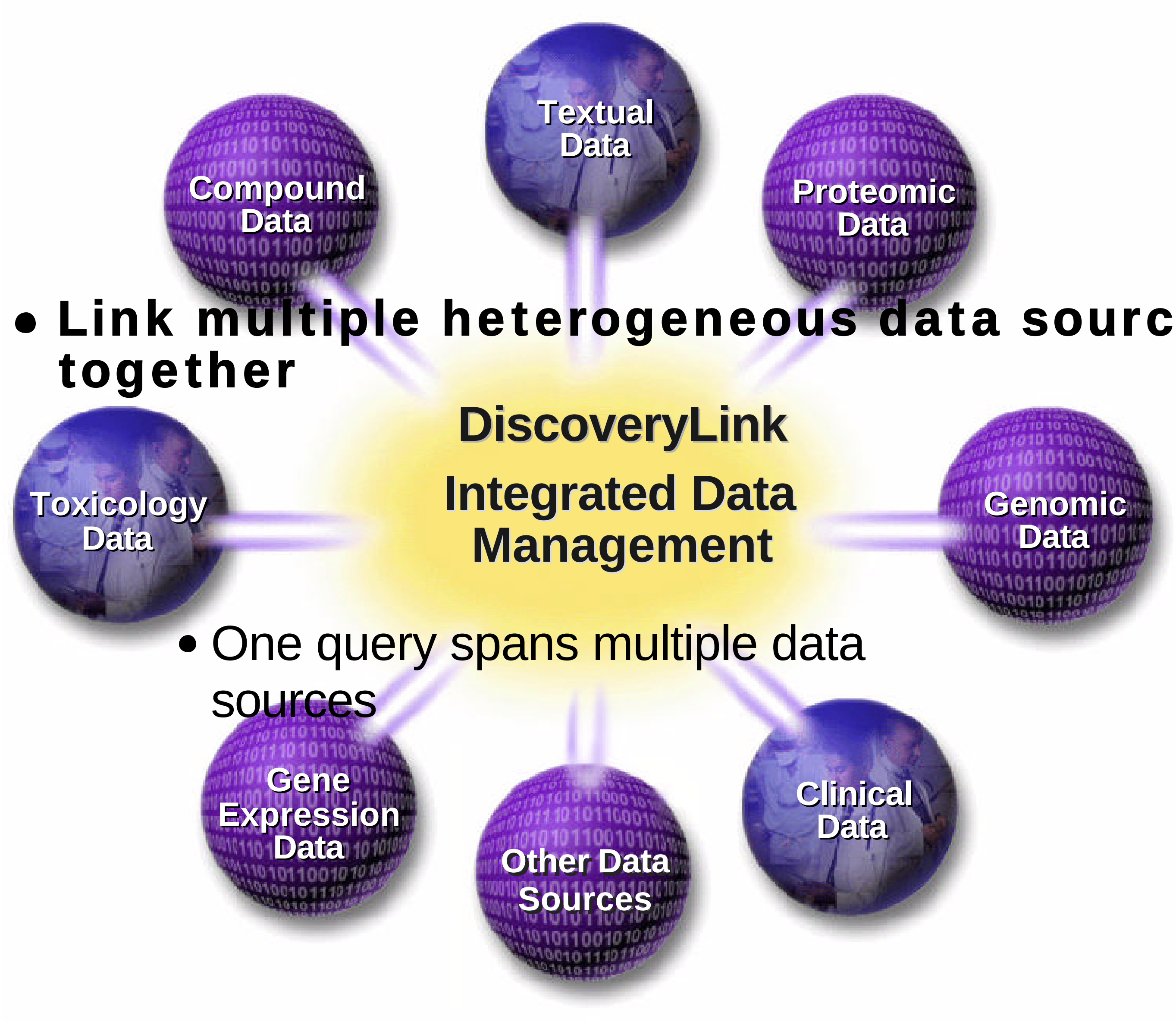


- ✓ Improved R&D productivitiy
- ✓ Efficiency in R&D process
- ✓ Competitve advantage in the marketplace
- ✓ Flexibility for future growth
- ✓ Lowered cost for scientific discovery



New technologies for Life Sciences

Integrated Data Management

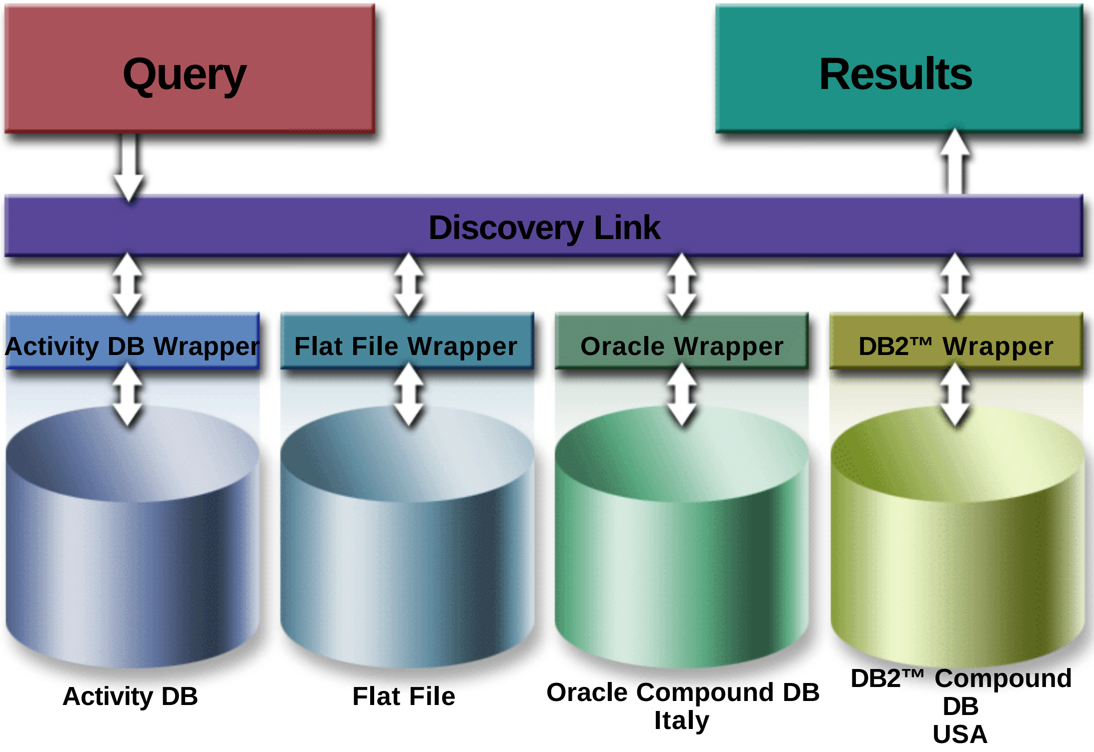


- **Link multiple heterogeneous data sources together**
- One query spans multiple data sources

DiscoveryLink

Solution: DiscoveryLink Enabling researchers to find critical needles in a haystack of data and documents

"Show me all the compounds similar to ketanserin that have been tested against members of the serotonin family and have the characteristics of a good drug."



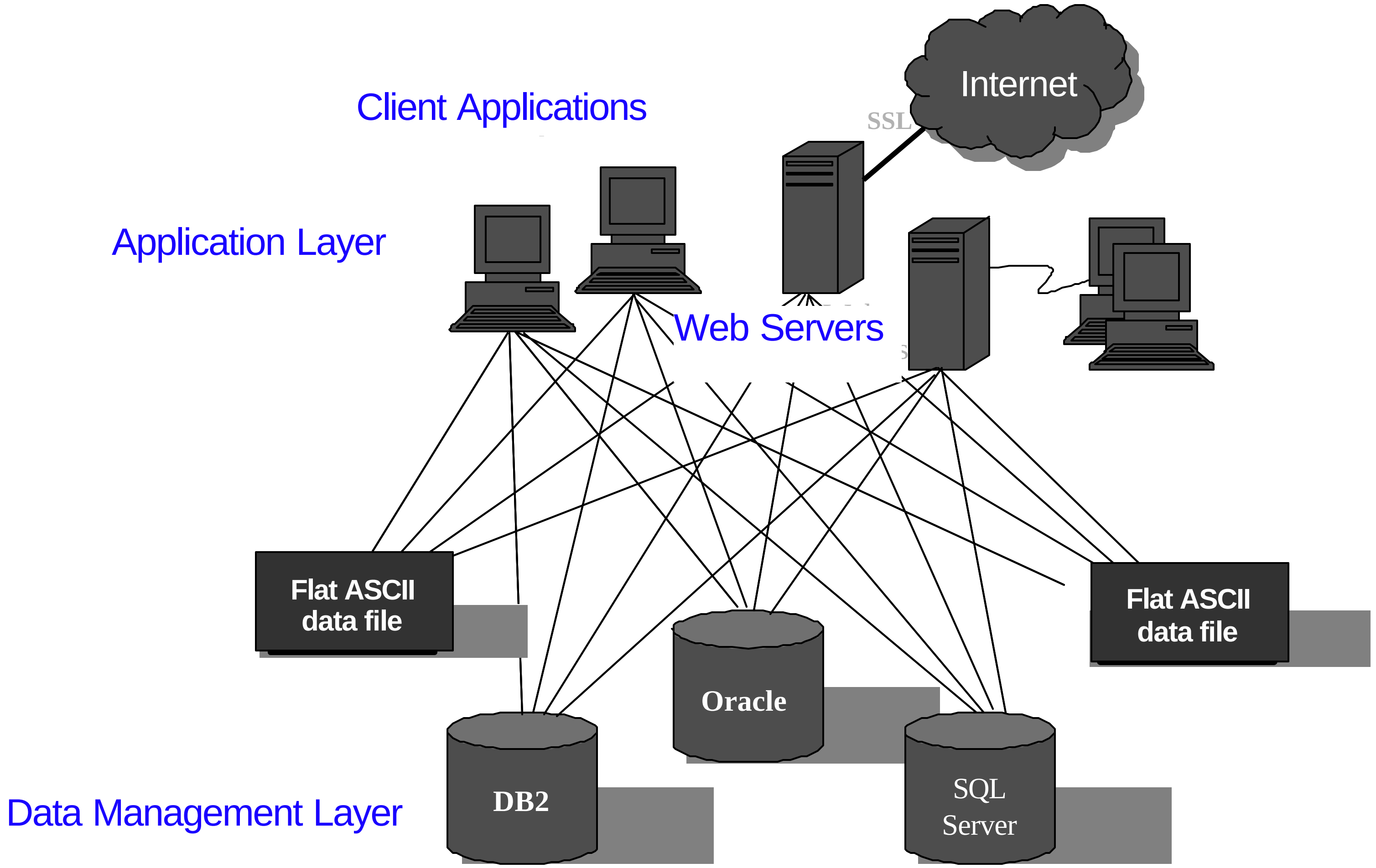
Capabilities:

- Accesses multiple and specialized databases with a single query
- Provides a single format virtual database view of multiple heterogeneous data sources
- Complements and extends existing data warehouse capabilities; eliminates the need to build query data warehouses
- Integrates analysis tools and business intelligence

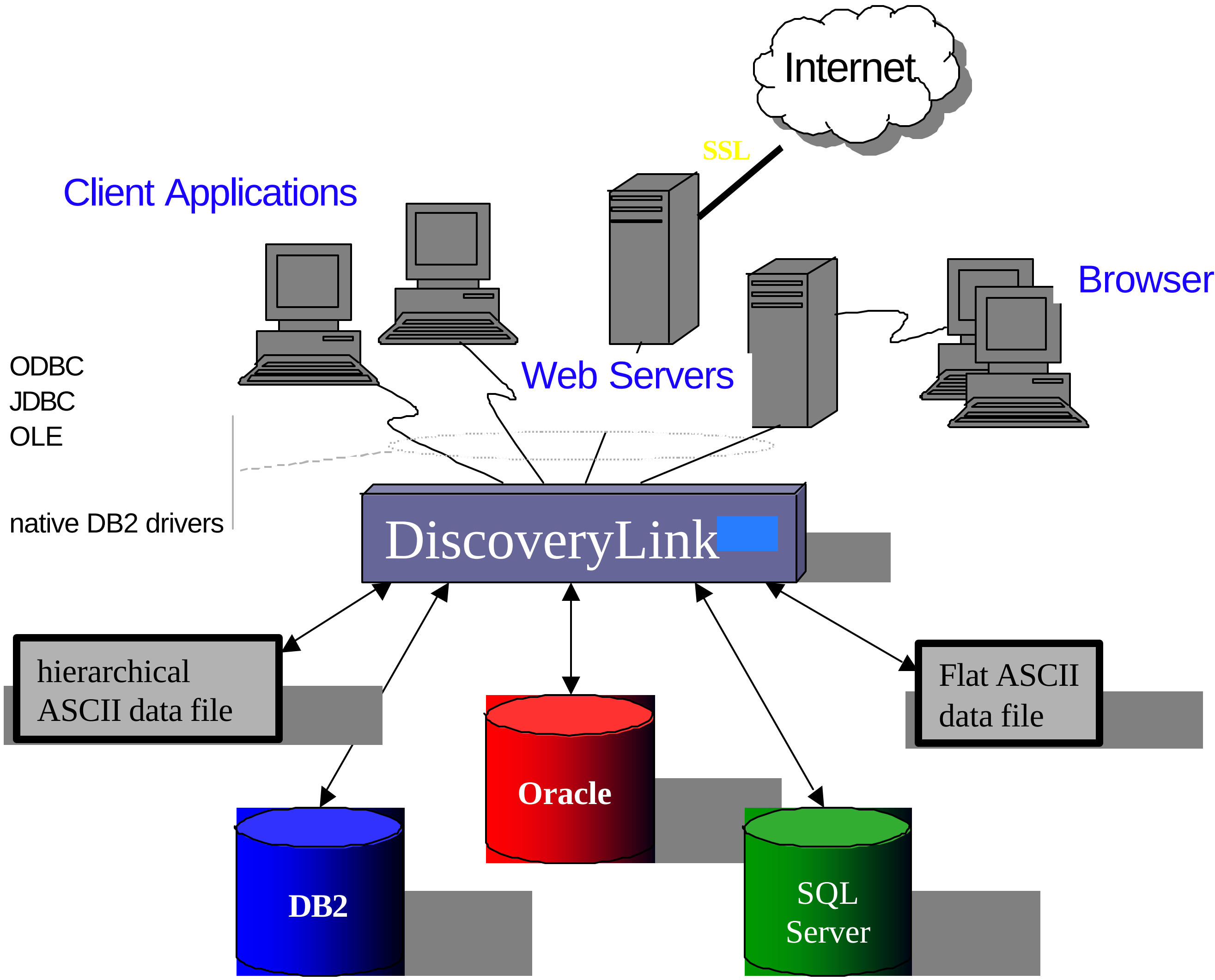
Benefits

- Provides a federated or single “virtual database” to applications
- Appears to be one data source
- Supports a high level query language (SQL)
- Integrate data from different data sources
- Diverse types of data
- Diverse sources
- One query can combine data from multiple sources
- No perturbation of existing data sources
- Exploit capabilities of existing sources
- To search for and manipulate data
- Lose no functionality

Without integration layer

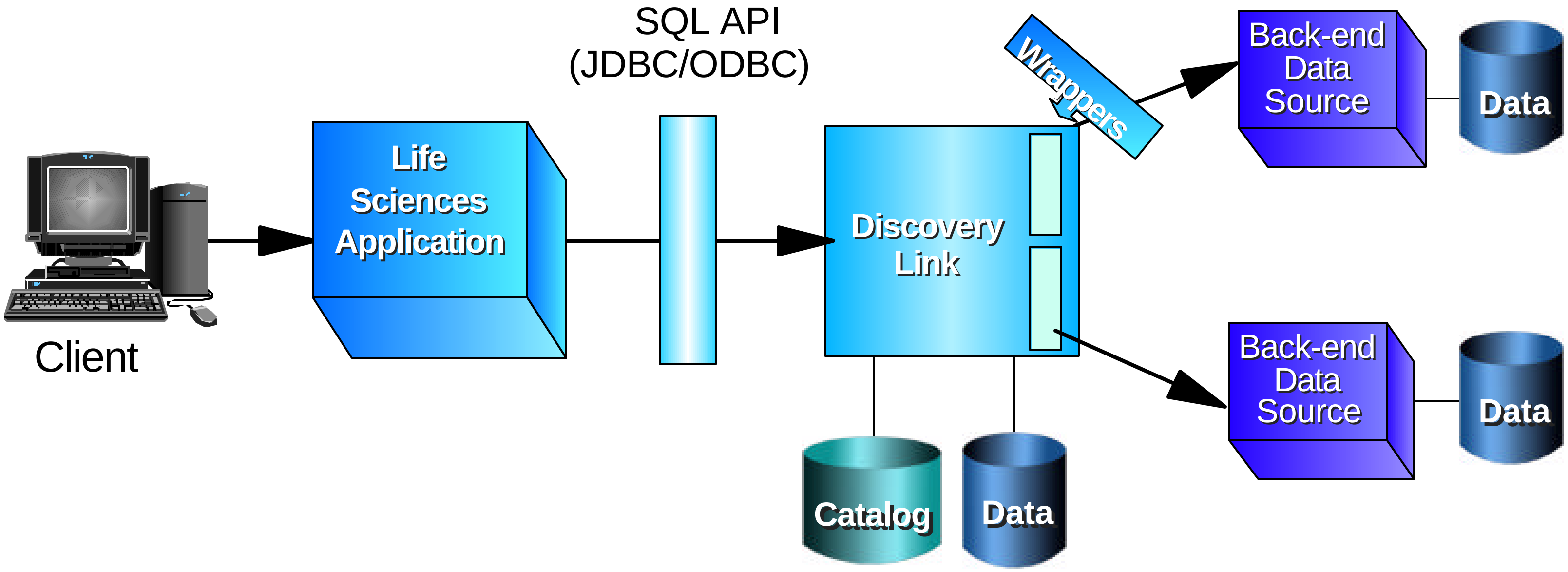


With integration layer



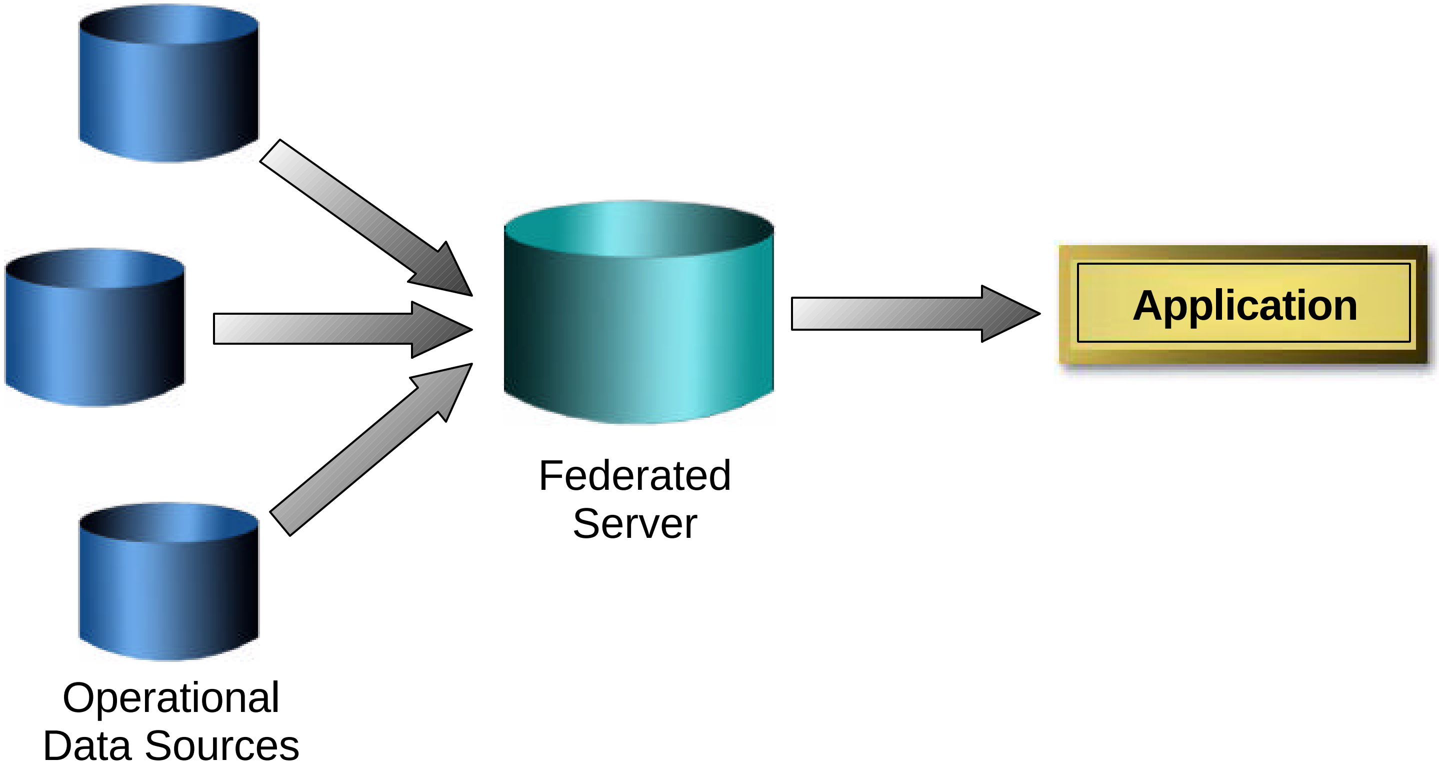
Architecture

- DiscoveryLink (DB2)® Federated Database Engine
  - DB2 drives DiscoveryLink **but it does not** replace existing client databases!
- Powerful **query processing** engine in **federated server**
- Logical decomposition and distribution of queries
- Cost-based optimizer to choose query plan



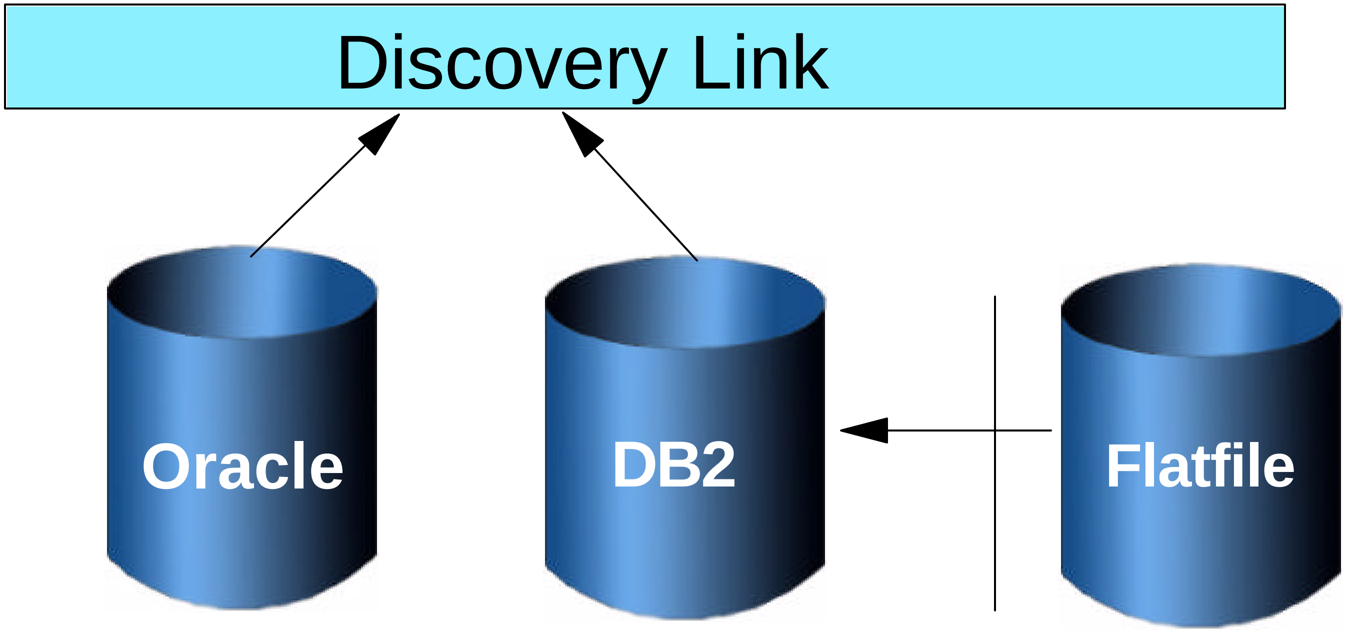
A Federated Database

- Data remains in the original separate sources
- All operational data sources accessible with a single query
- Query optimization on all data sources



Query 1

How similar is gene X to sequences within Genbank and within my in-house proprietary genome?



Discovery Link

Oracle

DB2

Flatfile

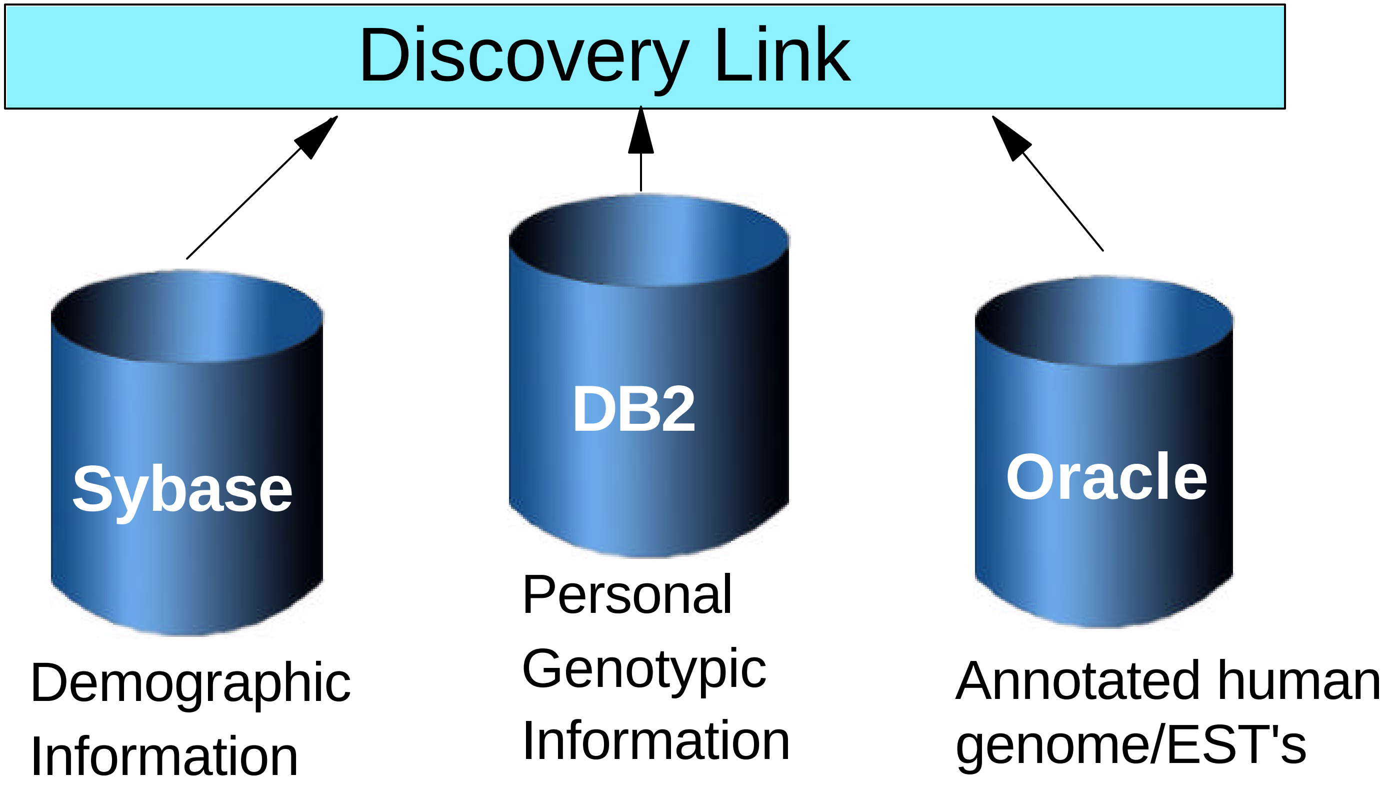
In-house Sequencing results

Downloaded version of Genbank

Genbank

Query 2

What gene or genes affect the reaction of some people to antibiotic X?



Discovery Link

Sybase

DB2

Oracle

Demographic Information

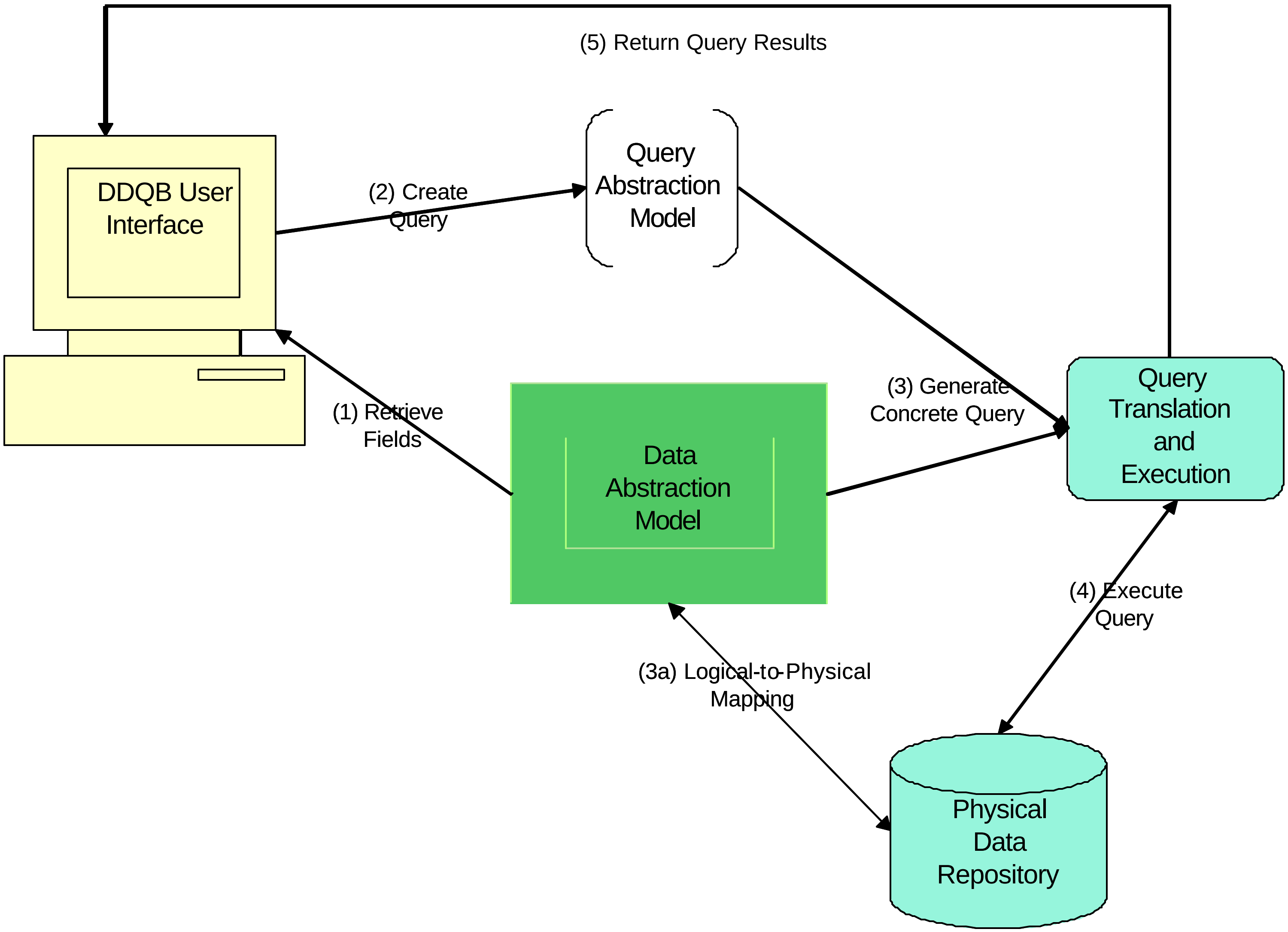
Personal Genotypic Information

Annotated human genome/EST's

What is DDQB?

- DDQB stands for:
  - Data Discovery Query Builder
- DDQB is a framework supporting description and execution of queries stated in abstract, implementation neutral terms
- Based on the concept of an abstract query
  - Queries stated in end user terms; not tied to a particular data representation, schema or location
  - Converted to a concrete query language like SQL for execution
  - Represented in XML
- Uses XML-based data abstraction model
  - Identifies logical fields referenced by abstract queries
  - Defines mapping to physical data representation
  - Supports 1-to-1 mapping between fields and physical data entities
  - Can also have logical fields that are:
    - Composed from 1 or more physical entities
    - Mapped to a subset of values for a given physical data entity
  - Can be statically defined or derived from other sources
- Includes a user interface used to create, execute and save abstract queries
  - Web-based UI
  - Can be extended via plugins for solution-unique behavior
  - Security, auditing, look and feel,…

DDQB Usage Flow



(5) Return Query Results

DDQB User Interface

(2) Create Query

Query Abstraction Model

(1) Retrieve Fields

Data Abstraction Model

(3) Generate Concrete Query

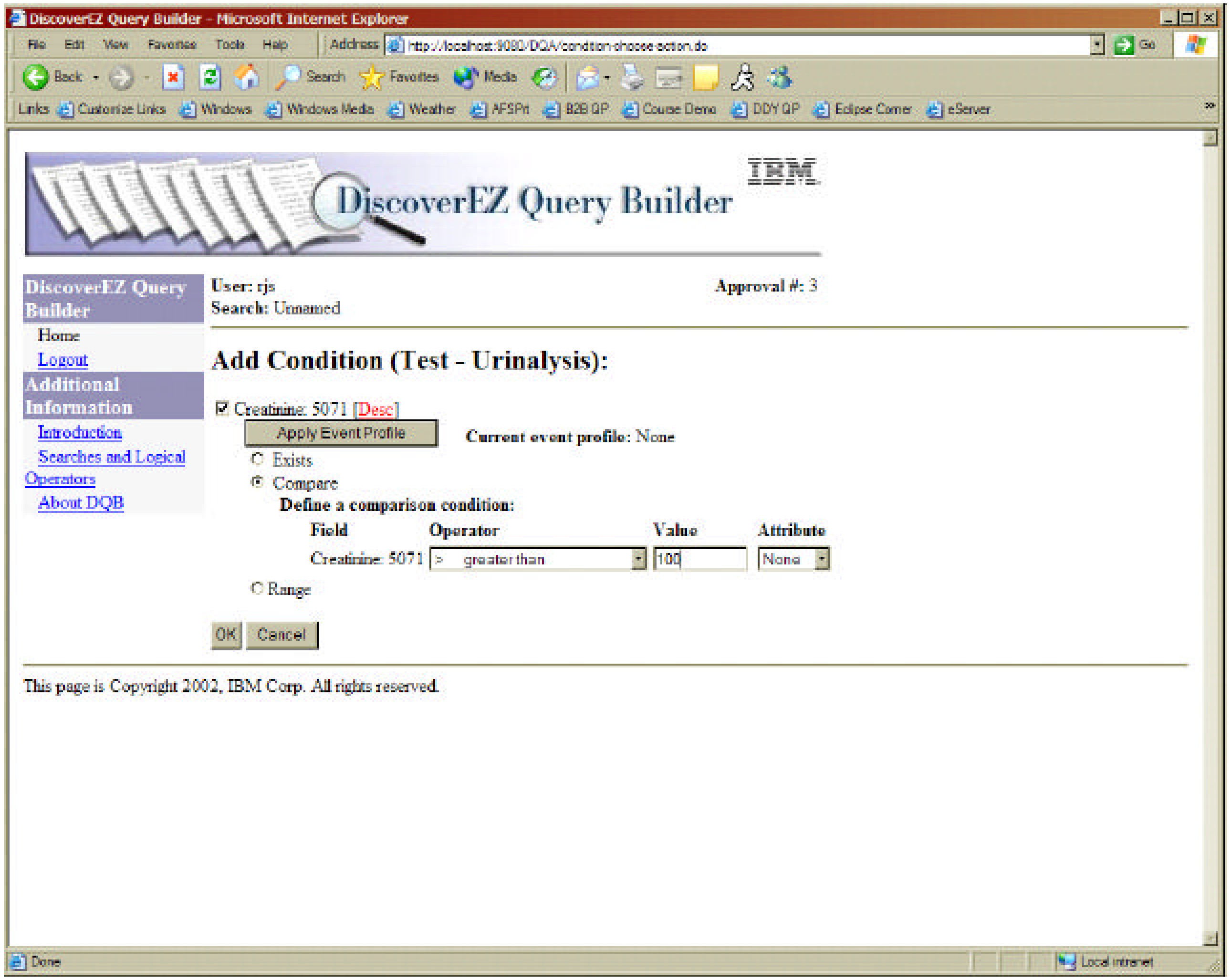
Query Translation and Execution

(4) Execute Query

(3a) Logical-to-Physical Mapping

Physical Data Repository

Specify Data Selection Criteria


DiscoverEZ Query Builder - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://localhost:9080/DQA/condition-choose-action.do
Go
Back Search Favorites Media
Links Customize Links Windows Windows Media Weather AFSPrt B2B QP Course Demo DDY QP Eclipse Corner eServer
IBM
DiscoverEZ Query Builder
Done
Local intranet

DiscoverEZ Query Builder

Home

Logout

Additional Information

Introduction

Searches and Logical Operators

About DQB

User: rjs **Approval #:** 3

Search: Unnamed

Add Condition (Test - Urinalysis):

☑ Creatinine: 5071 [Desc]

Apply Event Profile **Current event profile:** None

○ Exists

● Compare

Define a comparison condition:

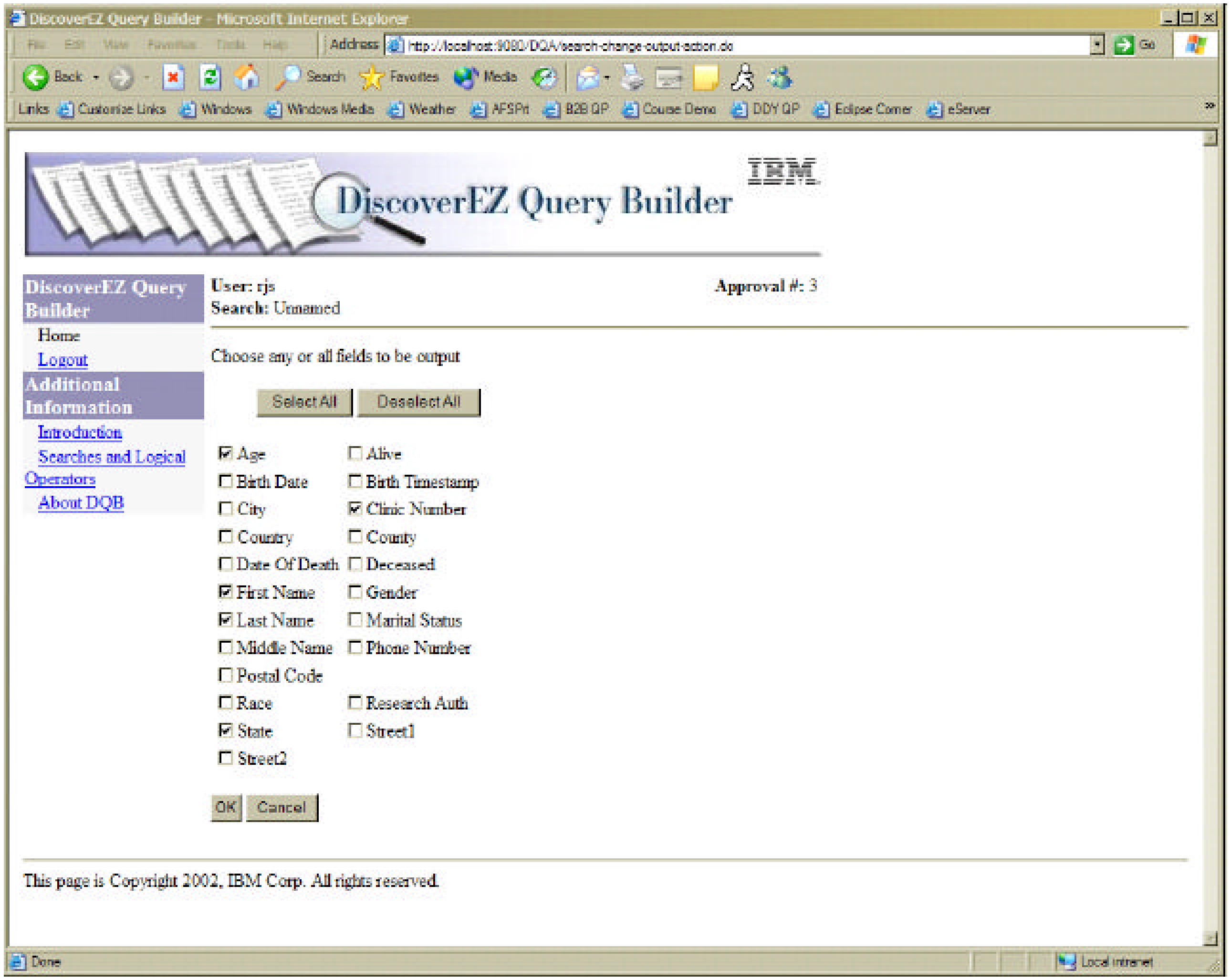
Field	Operator	Value	Attribute
Creatinine: 5071	> greater than	100	None

○ Range

OK Cancel

This page is Copyright 2002, IBM Corp. All rights reserved.

Select Query Output


DiscoverEZ Query Builder - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://localhost:9080/DQA/search-change-output-action.do
Go
Back Search Favorites Media
Links Customize Links Windows Windows Media Weather AFSPri B2B QP Course Demo DDY QP Eclipse Corner eServer
DiscoverEZ Query Builder
IBM
Done
Local intranet

DiscoverEZ Query Builder

Home

Logout

Additional Information

Introduction

Searches and Logical Operators

About DQB

User: rjs **Approval #:** 3

Search: Unnamed

Choose any or all fields to be output

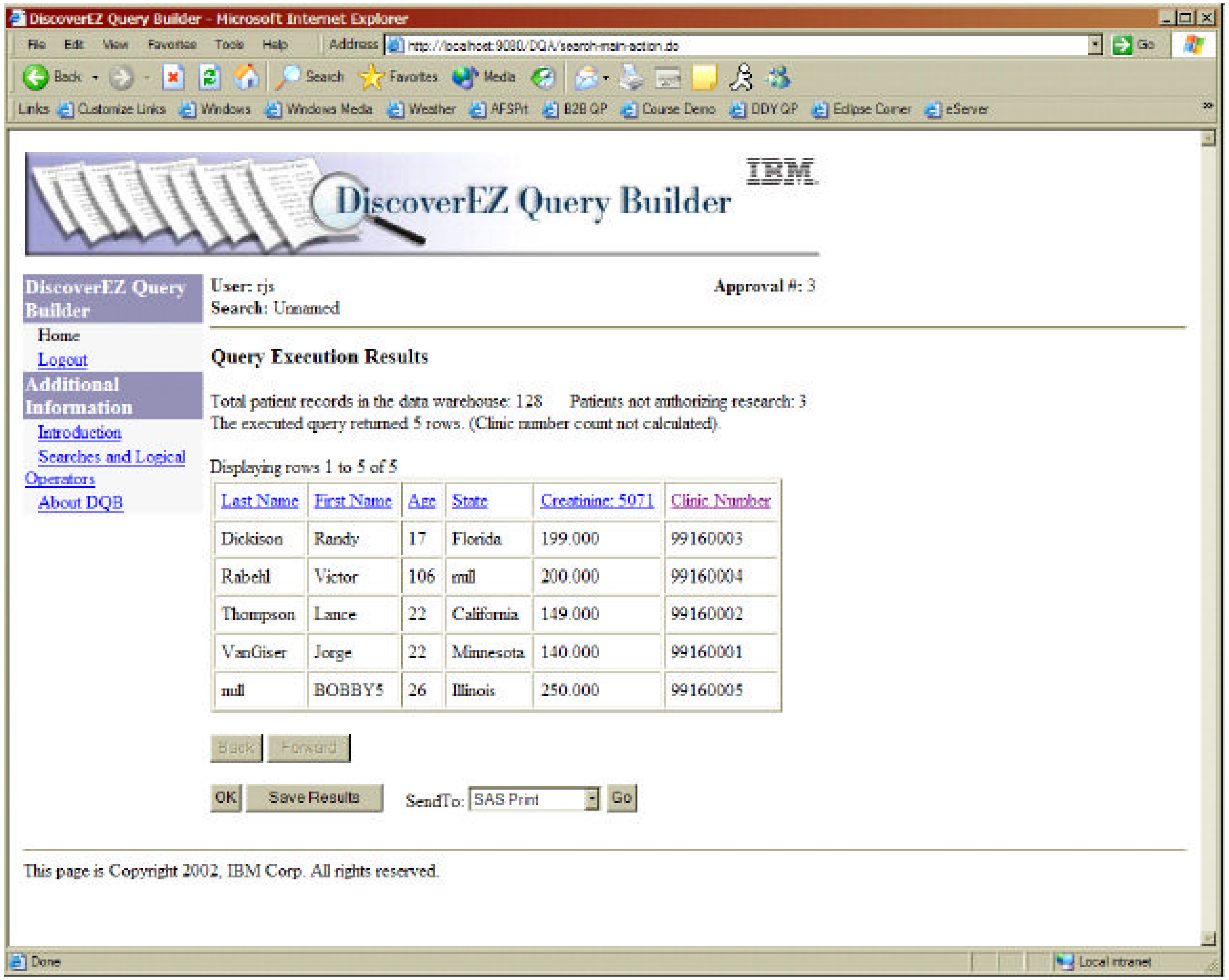
Select All Deselect All

☑ Age	☐ Alive
☐ Birth Date	☐ Birth Timestamp
☐ City	☑ Clinic Number
☐ Country	☐ County
☐ Date Of Death	☐ Deceased
☑ First Name	☐ Gender
☑ Last Name	☐ Marital Status
☐ Middle Name	☐ Phone Number
☐ Postal Code	
☐ Race	☐ Research Auth
☑ State	☐ Street1
☐ Street2	

OK Cancel

This page is Copyright 2002, IBM Corp. All rights reserved.

Execute, Display and Analyze Results


DiscoverEZ Query Builder - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://localhost:9080/DQA/search-main-action.do
Go
Back Search Favorites Media
Links Customize Links Windows Windows Media Weather AFSPit B2B QP Course Demo DDY QP Eclipse Corner eServer
DiscoverEZ Query Builder
IBM
Done
Local intranet

DiscoverEZ Query Builder

- Home
- Logout

Additional Information

- Introduction
- Searches and Logical Operators
- About DQB

User: rjs
Search: Unnamed

Approval #: 3

Query Execution Results

Total patient records in the data warehouse: 128 Patients not authorizing research: 3
The executed query returned 5 rows. (Clinic number count not calculated).

Displaying rows 1 to 5 of 5

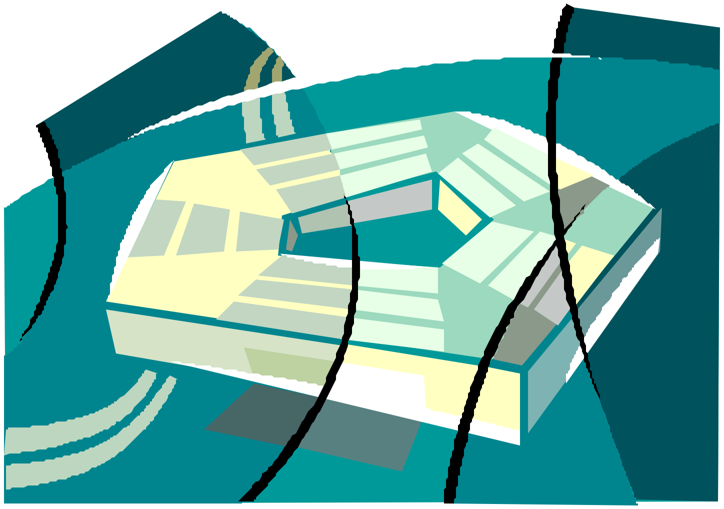
Last Name	First Name	Age	State	Creatinine: 5071	Clinic Number
Dickison	Randy	17	Florida	199.000	99160003
Rabehl	Victor	106	null	200.000	99160004
Thompson	Lance	22	California	149.000	99160002
VanGiser	Jorge	22	Minnesota	140.000	99160001
null	BOBBY5	26	Illinois	250.000	99160005

Back Forward

OK Save Results SendTo: SAS Print Go

This page is Copyright 2002, IBM Corp. All rights reserved.

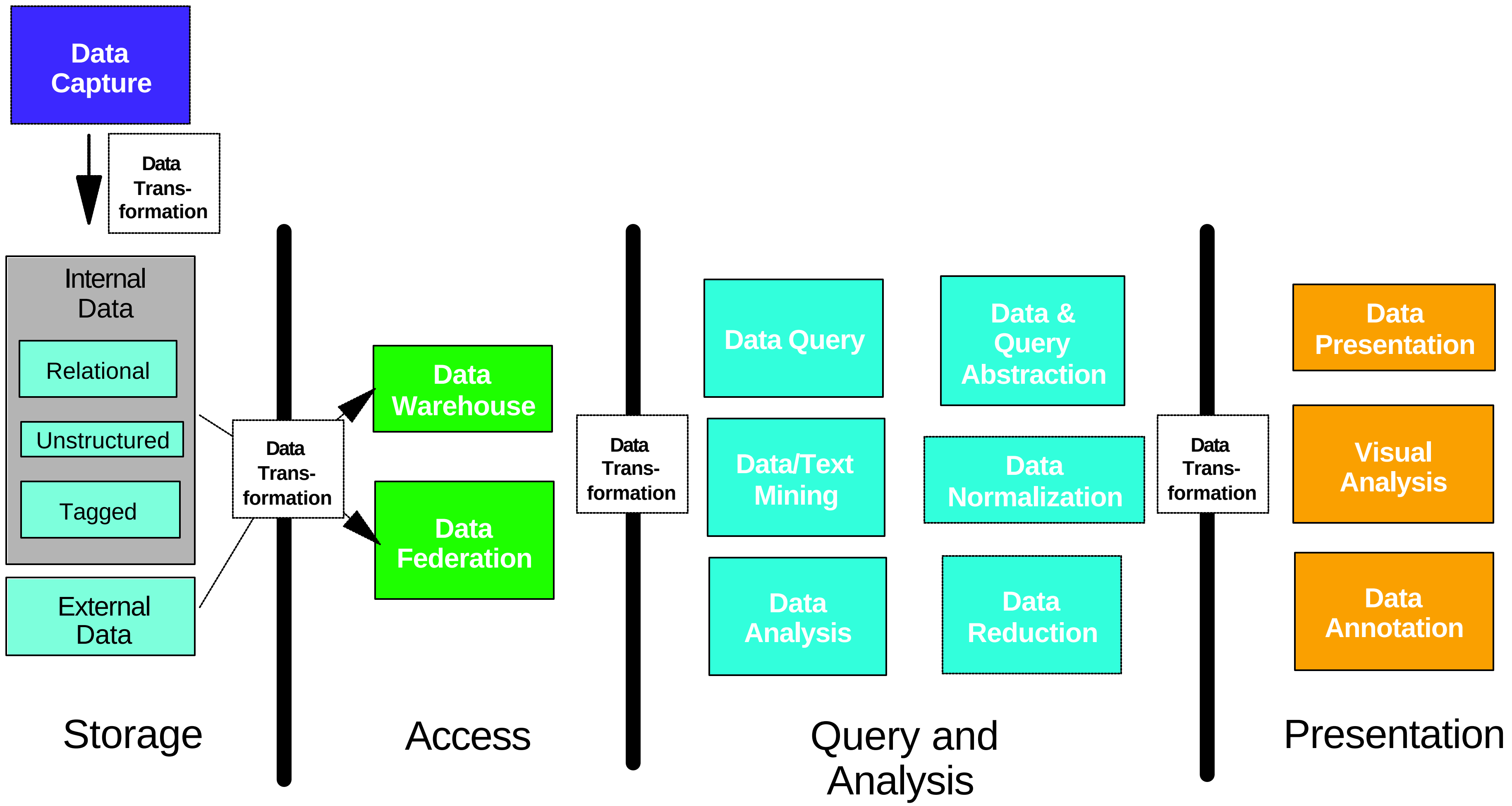
Life Sciences Framework Architecture



Understanding the customer's Data Management requirements

- **Every customer has existing data or needs to generate new data that needs to be manipulated and analyzed**
- **Understanding the data management requirements is key to developing a unique solution for the customer**

The Ideal Data Model

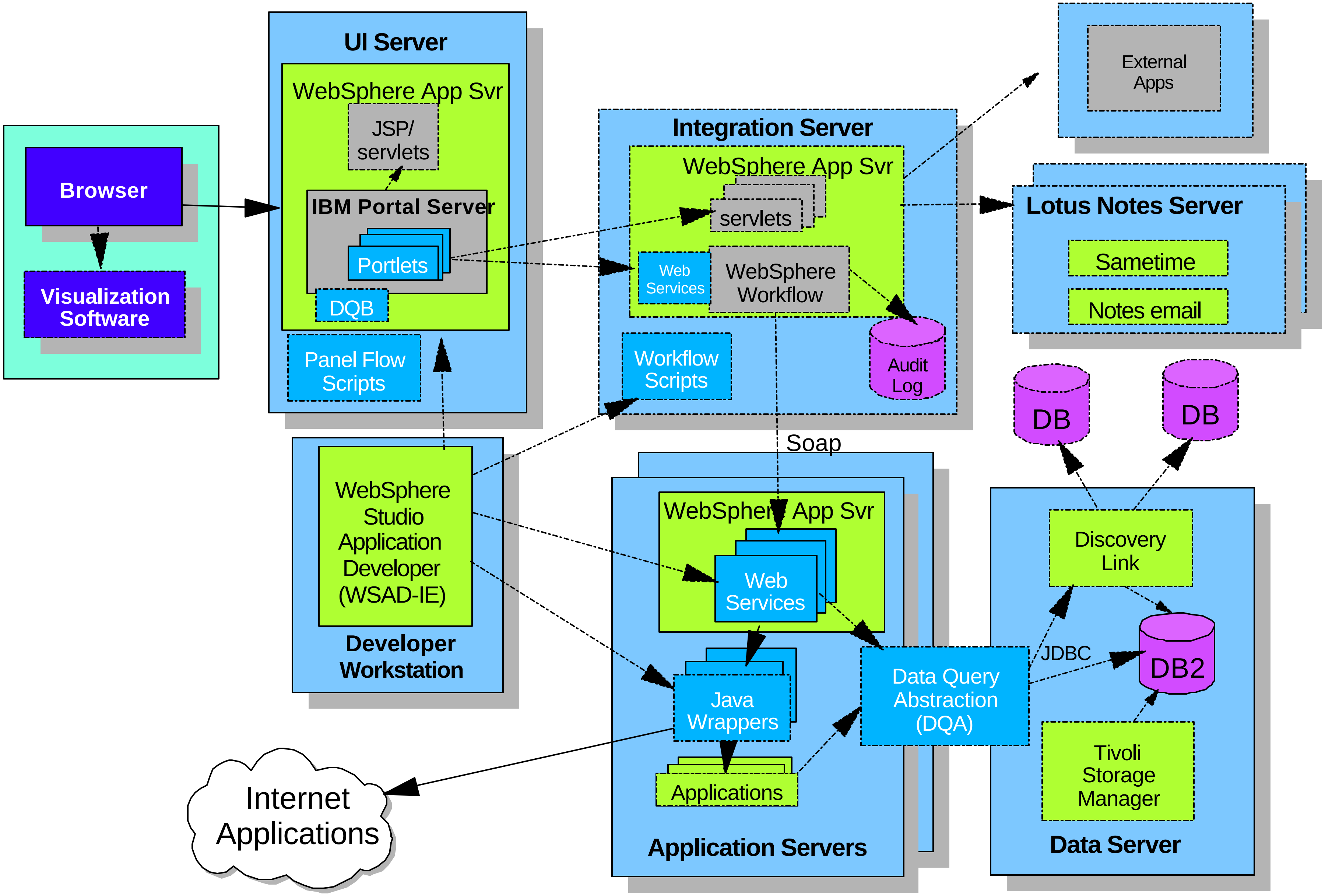


- Every instalation's data model is unique as is the dynamics of data motion and manipulation
- However they all perform the above types of operations

Understanding the customer's operational requirements

- Each customer solution has different characteristics that affect the preferred operational model
  - Price
  - Geography
  - Complexity
  - Existing infrastructure
  - Vendor bias

The Ideal Framework Operational Model



Preferred Fundamental Technologies

- **Java**
  - **Common cross-platform development language**
  - **Most of our tools assume Java as the base**
  - **Other languages supported through Java wrappers**
- **XML**
  - **Universal data interchange format**
  - **Self describing data is easier to transport**
  - **Internal data formats can be transformed to/from XML**
  - **There is likely a defined XML format for nearly every type of data**
  - **Numerous XML APIs**
    - **DOM (Document Object Model), SAX(Simple API for XML), JAXP(Java API for XML Parsing)**
  - **Numerous XML Manipulators**
    - **XSLT(Extensible Stylesheet Language for Transformations), XPath(XML Path Language)**

Preferred Fundamental Technologies

- **Web Services**
  - **Applications who's interface and binding can be defined via XML and can be accessed via XML-based messages over internet protocols**
  - **WSDL - Web Services Description Language**
    - **XML document that describes a web service (name, methods, arguments)**
  - **SOAP - Simple Object Access Protocol**
    - **Protocol for describing, via XML, a remote method to be invoked and returning the results as an XML document**
  - **UDDI - Universal Description, Discovery, and Integration**
    - **A registry of web services**

Preferred Fundamental Technologies

- **Composition**
  - **Workflow**
    - **WSFL - Web Services Flow Language**
      - **IBM proprietary, graph-oriented flow language**
      - **A version of WSFL is used by WSAD-IE**
    - **XLANG - XML Language**
      - **Microsoft proprietary, structure oriented flow language**
    - **BPELWS - Business Process Execution Language for Web Services**
      - **Merges WSFL and XLANG**
      - **Language for implementing a new web service as a composition of existing web services**
      - **Specification by BEA, IBM and Microsoft in initial public draft**
      - **BPWS4J engine available from alphaWorks**
      - **Incorporated into WSAD-IE in the future?**
      - **Defines an algorithm of steps (activities)**
      - **Primatives include <invoke>, <receive>, <reply>, <wait>, <assign>, <throw>, <terminate>, <empty>, <sequence>, <switch>, <while>, <pick>, <flow>**

Preferred Fundamental Technologies

- **Composition...**
  - **Business Process Integration**
    - **IBM CrossWorlds**
      - **Multi-threaded, Java based framework for collaborations**
      - **Automates transactions within a business process**
    - **MQSeries Workflow**
      - **Process deployment based on MQSeries**
    - **IBM Holosofx**
      - **Model and monitor business processes automated with MQSeries Worflow**

Preferred IBM Products

- **WebSphere**
  - **WAS**
    - Java-based Application Deployment Environment
    - Provides application services (transaction management, security, clustering, performance, availability, connectivity, scalability)
    - J2EE compliant
    - 5.0 is latest version
      - WAS, WAS Express, WAS Enterprise
  - **WebSphere Studio**
    - Java-based Application Development Environment
    - Runs on WebSphere Studio Workbench
      - IBM's version of the open source Eclipse platform
    - 5.0 is latest version
      - WSAD - WebSphere Studio Application Developer <---
        - Also WSAD-IE (Integration Editon) which provides workflow
      - WSSD - WebSphere Studio Site Developer
      - WSDD - WebSphere Studio Device Developer

Preferred IBM Products

- **WebSphere...**
  - **Portal**
    - **Single point of access to multiple types of information and applications**
    - **End user and administrator personalizatoin of portal views**
    - **Services: Single sign-on, security, content management, search, taxonomy, mobile devices, site analytics**
    - **4.1 is latest version**
      - **Portal for Multiplatforms**
      - **Portal Enable**
      - **Portal Extend**
      - **Portal Experience**
      - **Portal Express**

Preferred IBM Products

- **DB2**
  - **DB2 UDB V8.1 is latest version**
  - **DiscoveryLink**
    - **For Data Federation**
  - **Intelligent Miner for Data V8.1**
    - **Industrial strength mining technologies**
    - **Clustering, associations, sequential patterns, classification, prediction, similar time sequences**
  - **Intelligent Miner for Text**
    - **Advanced text mining and text search**
    - **Feature extraction, clustering, categorization, summarization**
- **DDQB**
  - **Data Discovery Query Builder**
  - **A framework supporting description and execution of queries stated in abstract, implementation neutral terms**

Preferred IBM Products

- Lotus
  - Notes and Domino
    - Messaging, collaboration, e-mail, calendaring, scheduling
  - Quickplace
    - Team collaboration of discussions, documents, tasks
  - Sametime
    - Chat, whiteboarding, application sharing
  - Lotus Knowledge Discovery Server
    - Search and expertise location solutions
    - Extracts, analyzes and categorizes structured and unstructured information
    - Generates Knowledge Maps for relevant content

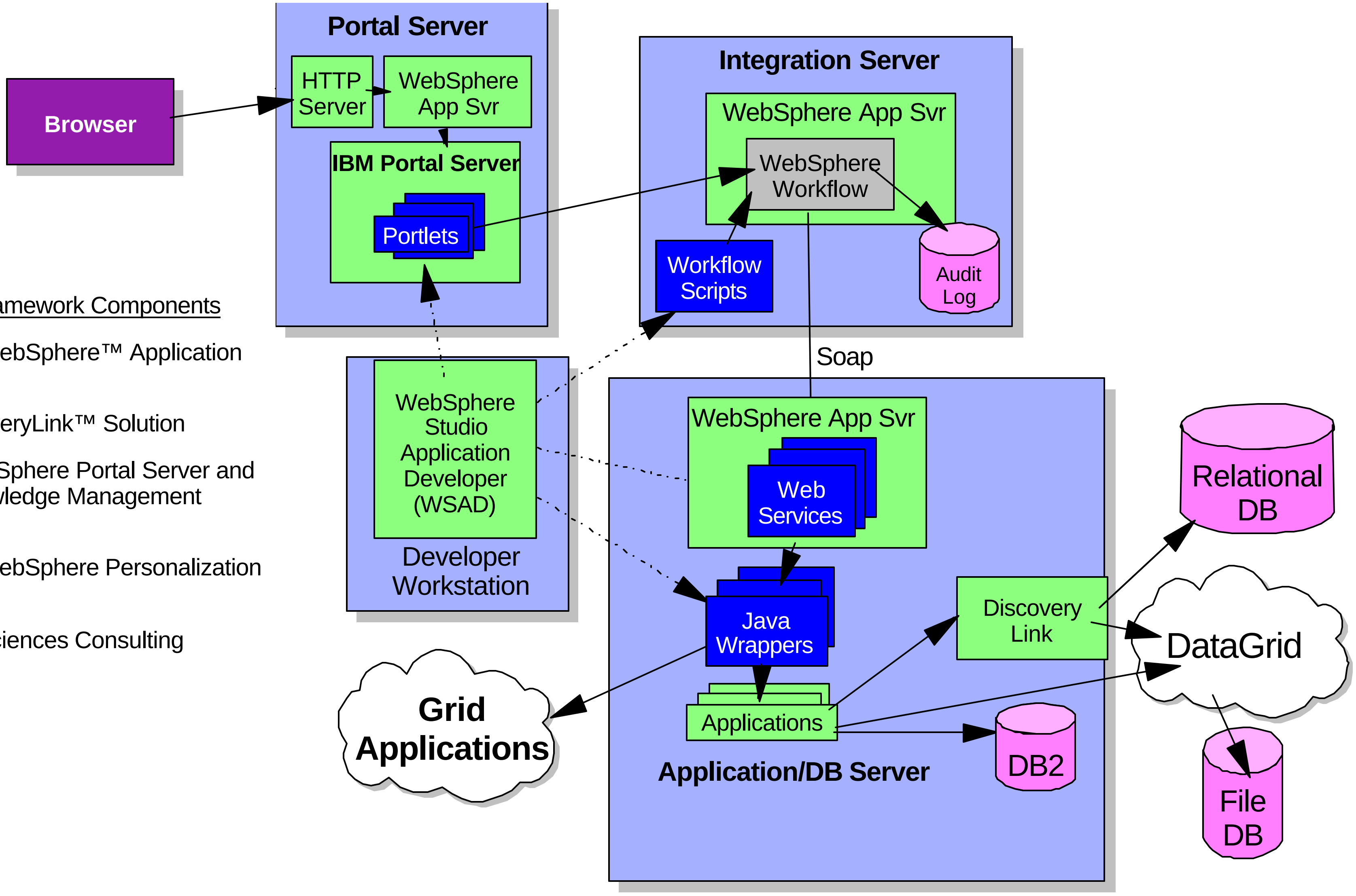
Preferred IBM Products

- Tivoli
  - Identity Manager
    - Centralized user account management
    - Self-service interfaces
  - Access Manager
    - Single sign-on, web administration, policy-based access control
  - Privacy Manager
    - Build, monitor and enforce privacy policies

Grid basics

- The End User
  - Can submit a job from an end system where neither the datasets nor the applications are installed.
  - Does not have to know where those datasets and applications actually reside
  - Can have the results stored on a local file system and can share that data according to individual policy
- The IT Administrator
  - Can install the datasets and applications once and thus manage a single copy
  - Can “scavenge” disk in a Linux cluster
  - Has the tools to implement policy across administrative domains

Grid Operational Model



Core IBM Framework Components

- The IBM WebSphere™ Application server
- IBM DiscoveryLink™ Solution
- IBM's WebSphere Portal Server and Lotus Knowledge Management Solutions
- The IBM WebSphere Personalization Server
- IBM Life Sciences Consulting

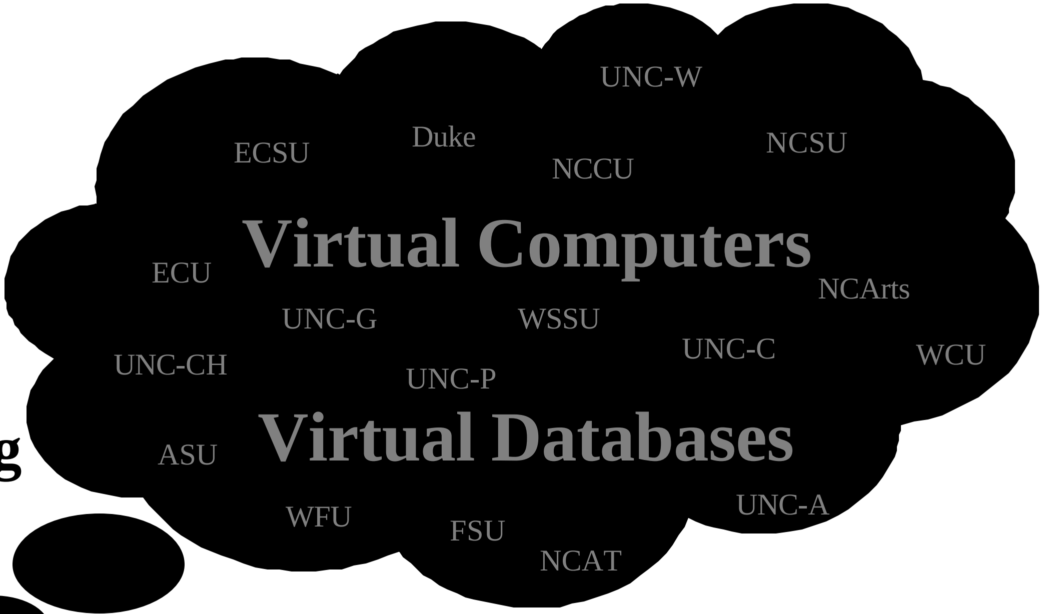
North Carolina Biogrid - Project Goals

- **Build a production infrastructure that:**
  - **Attracts Biotechnology Investment in the State of North Carolina**
  - **Serves a diverse community of researchers and educators**
  - **Virtualizes compute, storage, data, and network resources**
  - **Provides a unified view to a growing set of distributed resources**
  - **Scales with number of users and resource requirements**
  - **Embraces emerging technologies in distributed computing**
  - **Leverages our resources and our strengths**
  - **Allow the scientist to concentrate on science**
- **Allow the systems administrator to concentrate on IT**
- **Enable and facilitate innovation in life sciences research**
- **Built-in measurement capabilities for:**
  - **Measuring success**
  - **Capturing usage data**

NorthCarolina Biogrid

Attributes

- Single sign-on, security
- Policy-based resource sharing



UNC-W

ECSU Duke NCCU NCSU

Virtual Computers

ECU NCArts

UNC-G WSSU

UNC-CH UNC-C WCU

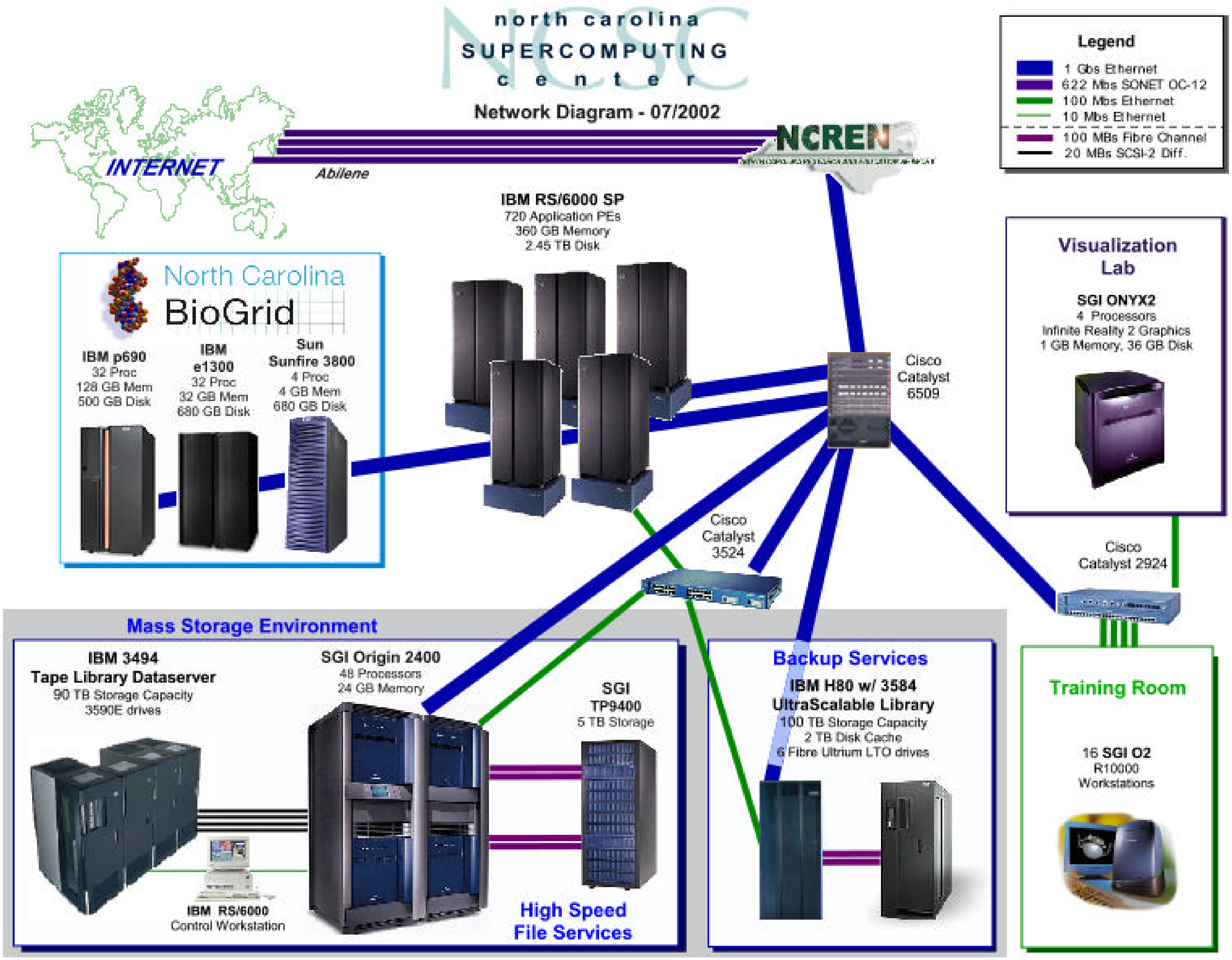
UNC-P

ASU Virtual Databases

WFU FSU UNC-A

NCAT

- Unified view of data and computers
  - Computers and data appear to be local
- Efficient access to large data sets
  - Caching
  - Replication



north carolina SUPERCOMPUTING center

Network Diagram - 07/2002

Legend

🟦 1 Gbs Ethernet
🟪 622 Mbs SONET OC-12
🟩 100 Mbs Ethernet
— 10 Mbs Ethernet
🟪 100 MBs Fibre Channel
— 20 MBs SCSI-2 Diff.

INTERNET

Abilene

NCREN

IBM RS/6000 SP
720 Application PEs
360 GB Memory
2.45 TB Disk

North Carolina BioGrid

IBM p690
32 Proc
128 GB Mem
500 GB Disk

IBM e1300
32 Proc
32 GB Mem
680 GB Disk

Sun Sunfire 3800
4 Proc
4 GB Mem
680 GB Disk

Cisco Catalyst 6509

Visualization Lab

SGI ONYX2
4 Processors
Infinite Reality 2 Graphics
1 GB Memory, 36 GB Disk

Cisco Catalyst 3524

Cisco Catalyst 2924

Mass Storage Environment

IBM 3494
Tape Library Dataserver
90 TB Storage Capacity
3590E drives

SGI Origin 2400
48 Processors
24 GB Memory

SGI TP9400
5 TB Storage

IBM RS/6000
Control Workstation

High Speed File Services

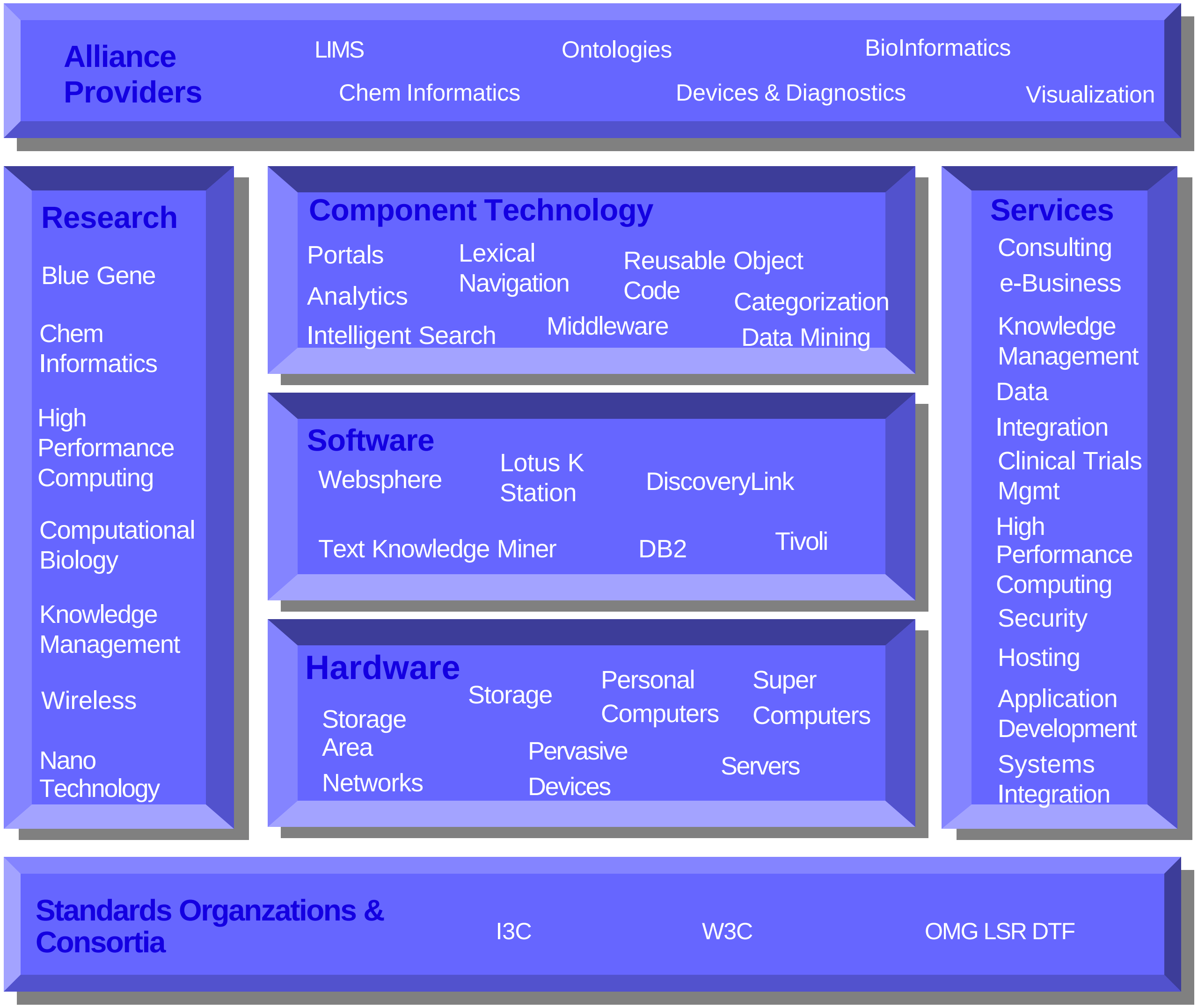
Backup Services

IBM H80 w/ 3584
UltraScalable Library
100 TB Storage Capacity
2 TB Disk Cache
6 Fibre Ultrium LTO drives

Training Room

16 SGI O2
R10000
Workstations

Delivering end-to-end solutions for the Life Sciences industry



Alliance Providers

LIMS
Ontologies
BioInformatics
Chem Informatics
Devices & Diagnostics
Visualization

Research

Blue Gene
Chem Informatics
High Performance Computing
Computational Biology
Knowledge Management
Wireless
Nano Technology

Component Technology

Portals
Lexical Navigation
Reusable Object Code
Analytics
Categorization
Intelligent Search
Middleware
Data Mining

Software

Websphere
Lotus K Station
DiscoveryLink
Text Knowledge Miner
DB2
Tivoli

Hardware

Storage
Personal Computers
Super Computers
Storage Area Networks
Pervasive Devices
Servers

Services

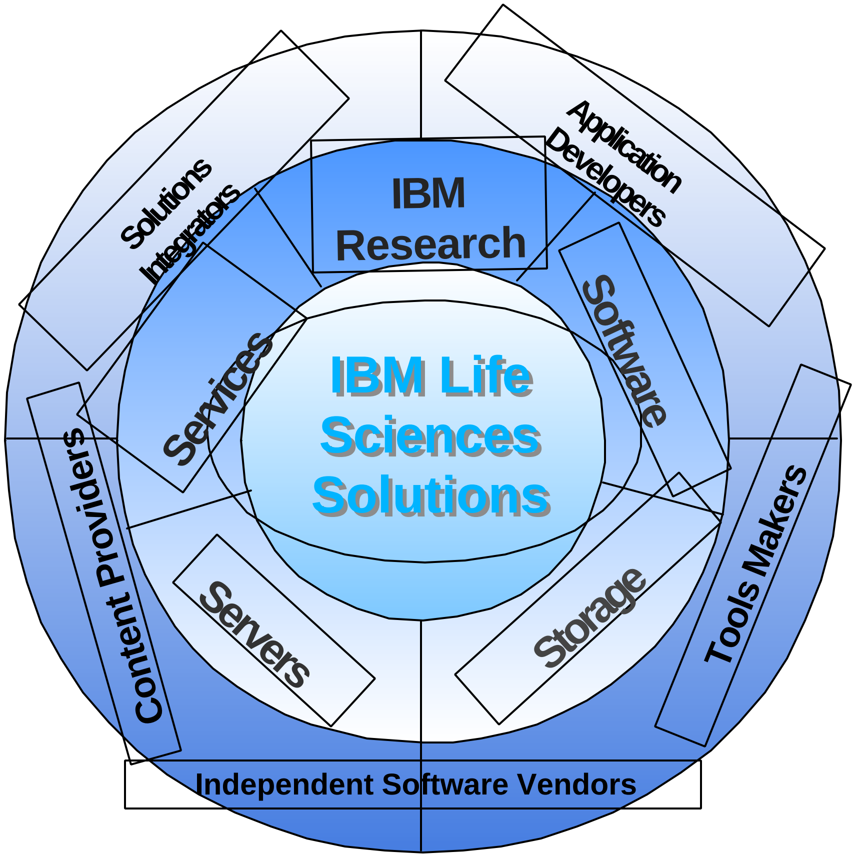
Consulting
e-Business
Knowledge Management
Data Integration
Clinical Trials Mgmt
High Performance Computing
Security
Hosting
Application Development
Systems Integration

Standards Organzations & Consortia

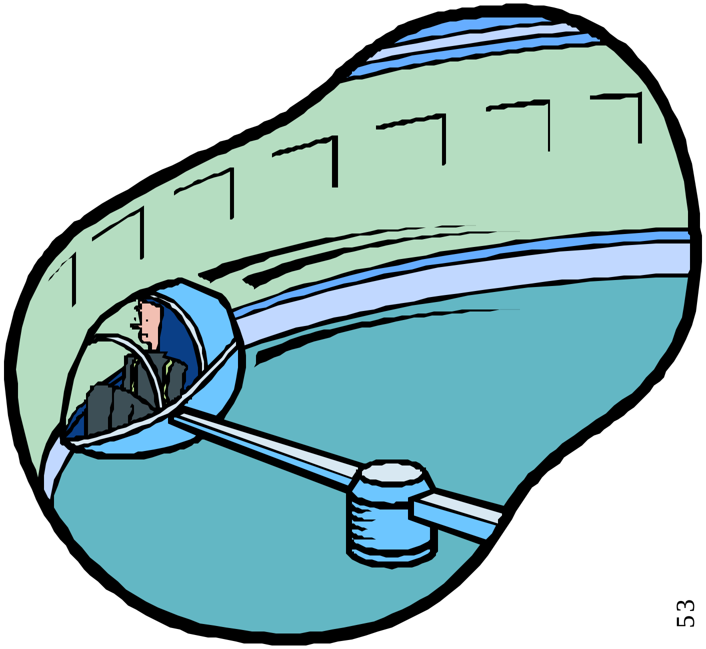
I3C
W3C
OMG LSR DTF

Leveraging and continuously building on IBM's capabilties

- **Research focused on Life Sciences Issues (over 50 Ph.Ds)**
  - IBM Computational Biology Center, IBM Deep Computing Institute
- **Dedicated Industry Business Unit**
  - Executive, Marketing and Sales teams, most with Life Sciences education or experience
  - Solution Development team with extensive IT and/or domain expertise
  - Longterm customer and partner relationships
- **Dedicated Global Consulting Units**
  - Life Sciences practice focused on R&D in pharmaceuticals and biotechs
  - Healthcare practice focused on Delivery in pharmaceuticals and point-of-care providers
- **Proven Technologies, Solutions, and Methodologies**



Pilot Engagements



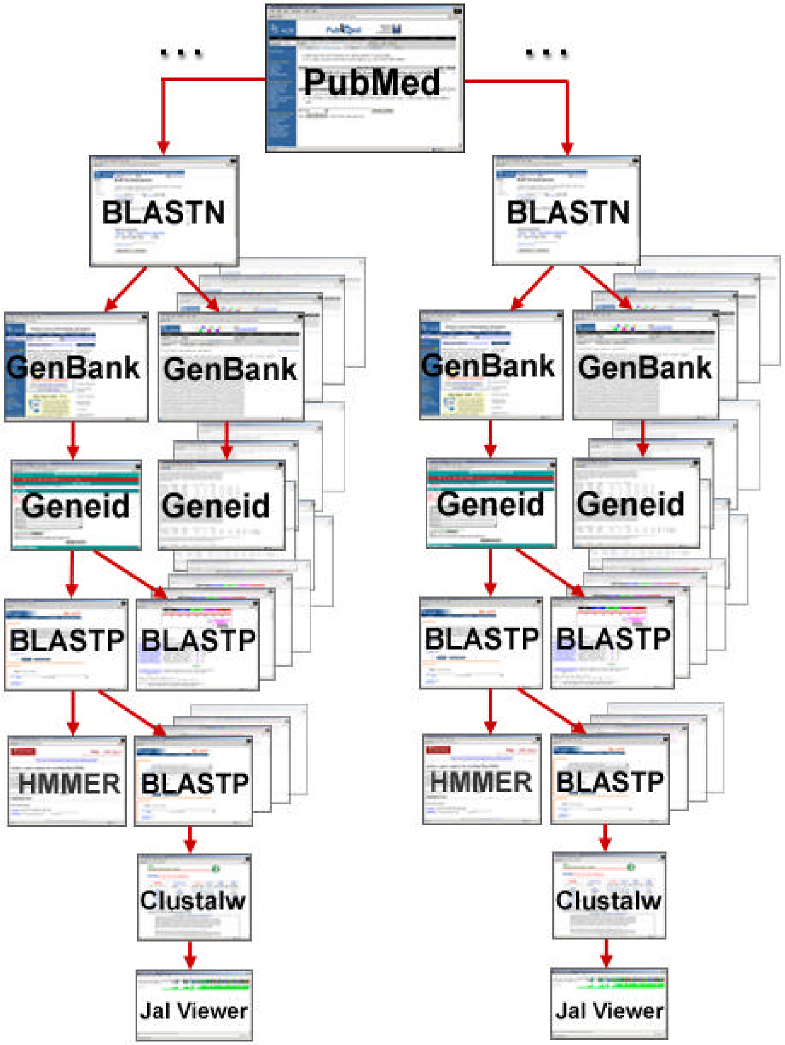
Pilot engagements: Strategic objectives

- **Validate the LS Framework model in customer environment**
- **Understand the customer requirements to improve our offerings and fill in the gaps**
- **Develop IBM software extensions and reusable assets for the LS industry**
- **Identify and recruit important business partners**
- **Develop solutions to solve business problems across the industry**
- **Establish IBM presence in major influencers (e.g leading research and medical institutions, universities)**

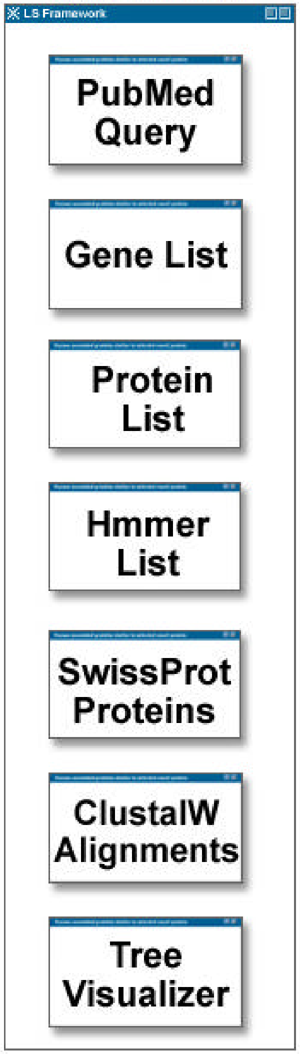
Scenario

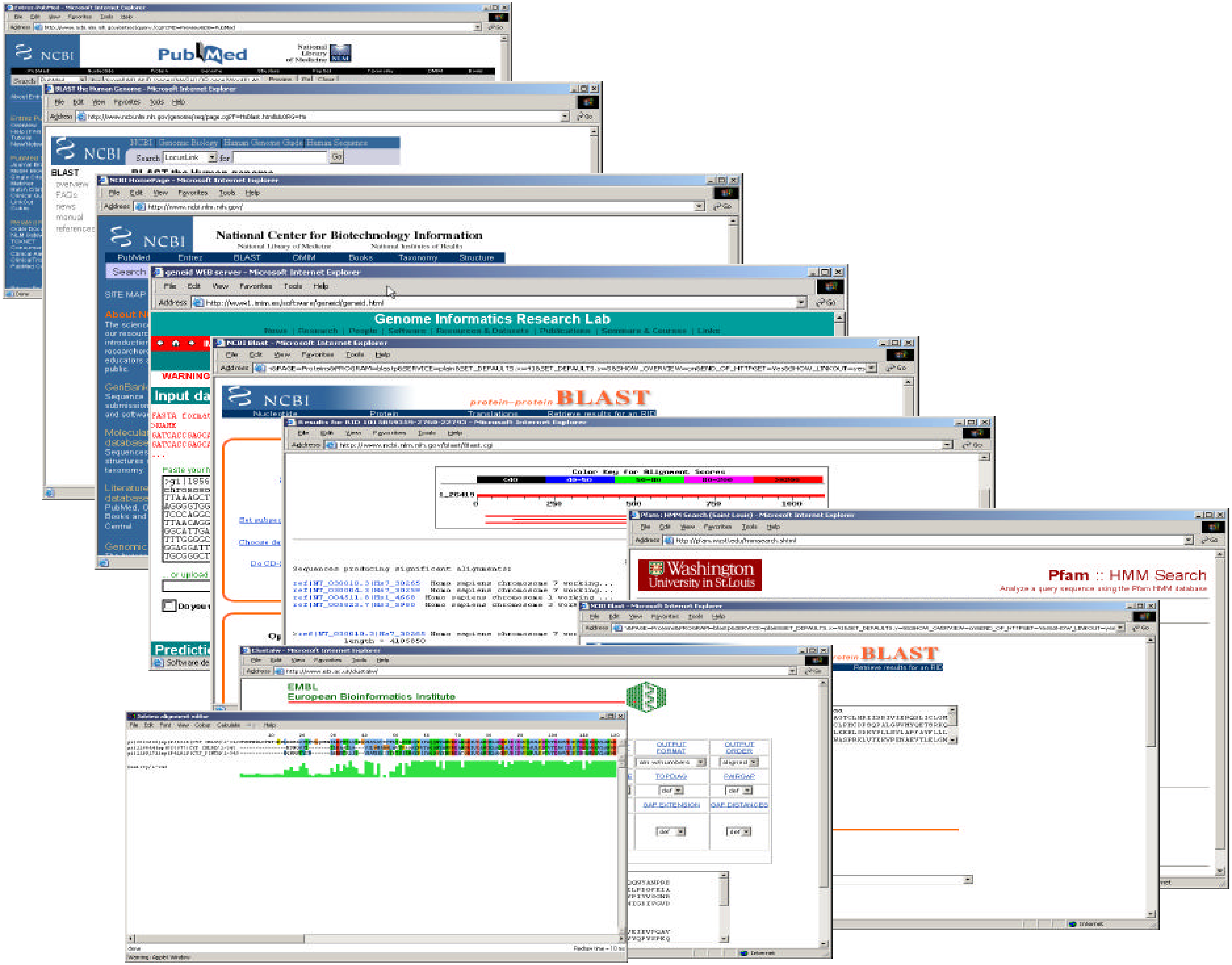
Challenge: Find & Characterize novel cancer related genes in genomic sequences

Screens Before using the LS Framework



Screens After using the LS Framework

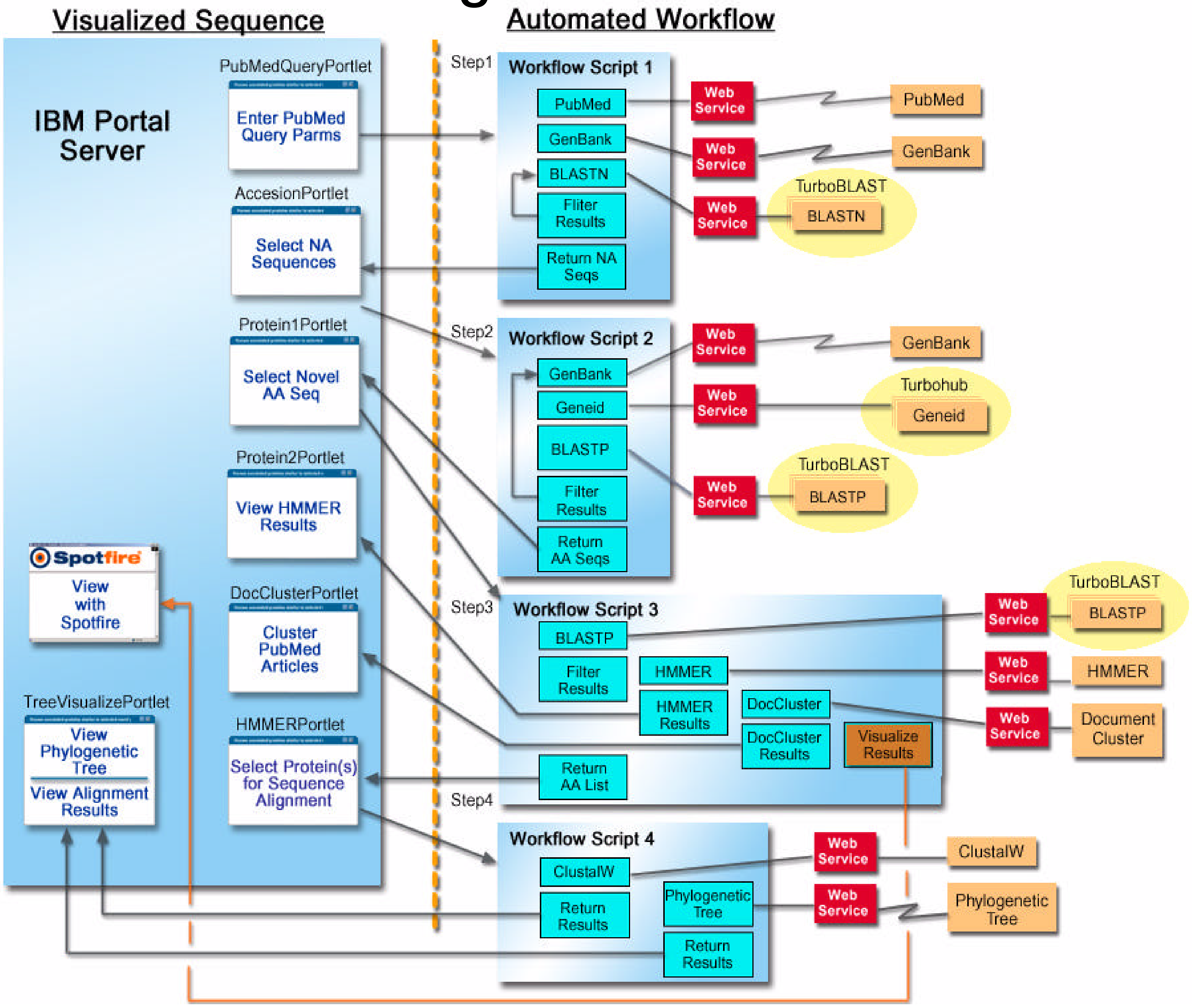



PubMed
NCBI
BLAST the Human Genome - Microsoft Internet Explorer
NCBI HomePage - Microsoft Internet Explorer
National Center for Biotechnology Information
PubMed Entrez BLAST OMIM Books Taxonomy Structure
geneid WEB server - Microsoft Internet Explorer
Genome Informatics Research Lab
NCBI Blast - Microsoft Internet Explorer
protein-protein BLAST
Color Key for Alignment Scores
Sequences producing significant alignments:
Pfam: HMM Search (Saint Louis) - Microsoft Internet Explorer
Washington University in St.Louis
Pfam :: HMM Search
ClustalW - Microsoft Internet Explorer
EMBL
European Bioinformatics Institute
Jalview alignment editor

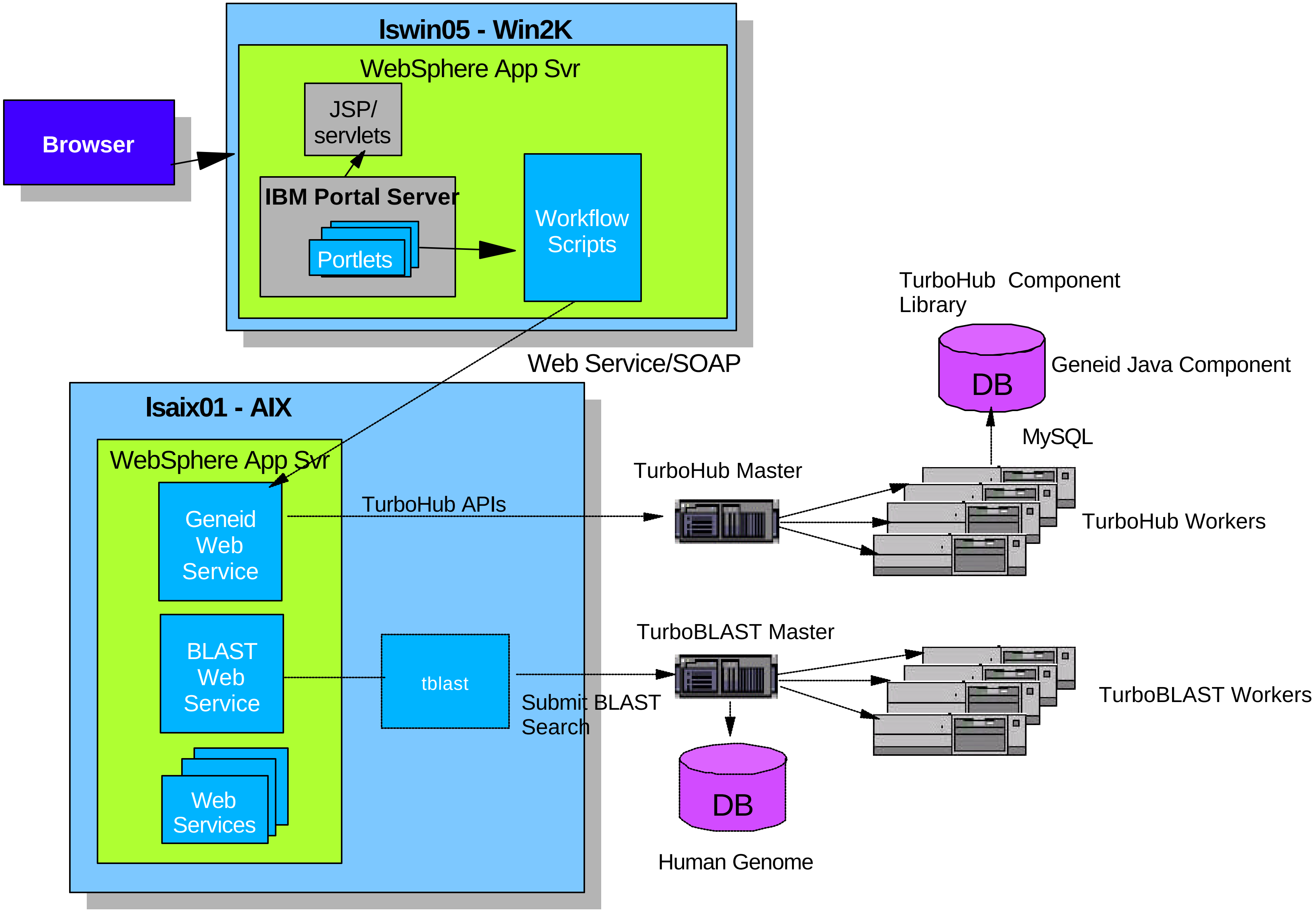
Framework Approach

- Build <u>Web Services</u> wrappers around the applications used by the researcher in this scenario
  - ► Some of these applications will be run locally
  - ► Some will be accessed via the Internet
- Automate the choreography of the applications through <u>workflow</u> scripts
- Provide user interaction through IBM's <u>Portal Server</u> interface
- Provide open infrastructure that integrates major Life Sciences applications and IBM research technologies

Novel Gene Finding Demo Overview



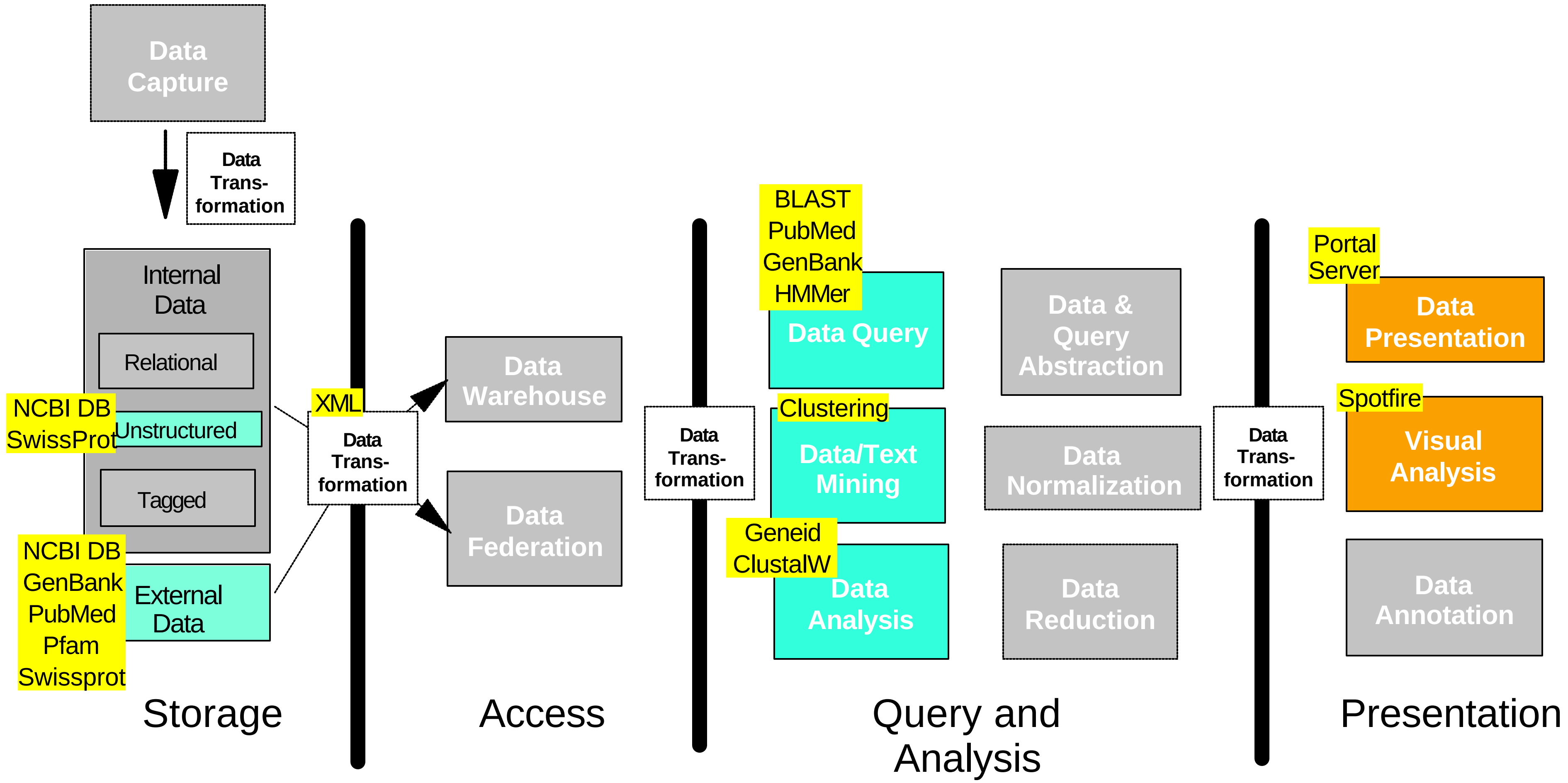
TurboBLAST and Geneid TurboHub System View



IBM Life Sciences

Novel Gene Finding Demo Overview

Find and characterize novel cancer-related genes in genomic sequences.



Data Capture

Data Trans-formation

Internal Data

Relational

Unstructured

Tagged

NCBI DB
SwissProt

External Data

NCBI DB
GenBank
PubMed
Pfam
Swissprot

XML

Data Trans-formation

Data Warehouse

Data Federation

Data Trans-formation

BLAST
PubMed
GenBank
HMMer

Data Query

Clustering

Data/Text Mining

Geneid
ClustalW

Data Analysis

Data & Query Abstraction

Data Normalization

Data Reduction

Data Trans-formation

Portal Server

Data Presentation

Spotfire

Visual Analysis

Data Annotation

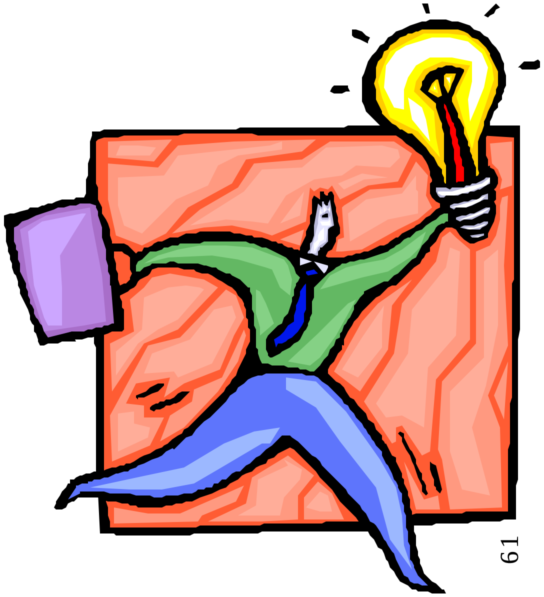
Storage

Access

Query and Analysis

Presentation

What have we learned?



So, What have we learned?

- **From a conceptual level, all customers have the same problem**
  - **Lots of data in various formats**
  - **Need to query and analyze the data**
  - **Need numerous ways to view the data**
  - **Which drives need to integrate multiple applications from various vendors**
- **However, each customer solution is unique**
  - **Different data, skills, expense structure, performance requirements, security requirements, etc...**

So, What have we learned?

- The Life Sciences Framework concept and technologies seems to address these problems quite well
  - Total end-to-end coverage
  - Flexibility to substitute various technologies
- We're getting tremendous re-use out of the assets we've developed so far
  - Keeping the assets single purpose and then chaining them together with workflow seems to be the right model
  - DDQB has been used in nearly all our engagements with rave reviews from the customers

So, What have we learned?

- The functionality of WebSphere Application Server, WebSphere Portal Server, and WebSphere Workflow are appreciated by the customers
  - Only a fraction of this functionality is actually used by any one customer
  - The advantages of Portal Server are hard to convey until a prototype is built