# Big Data Training Courses
# In Statistics Training Institute of Statistics Korea

Statistics Training Institute
Jeongran Kim and Jaemin Na

# Big Data Training Course of STI

1. Background
2. Big Data Program
3. Outcome

# Future Plans

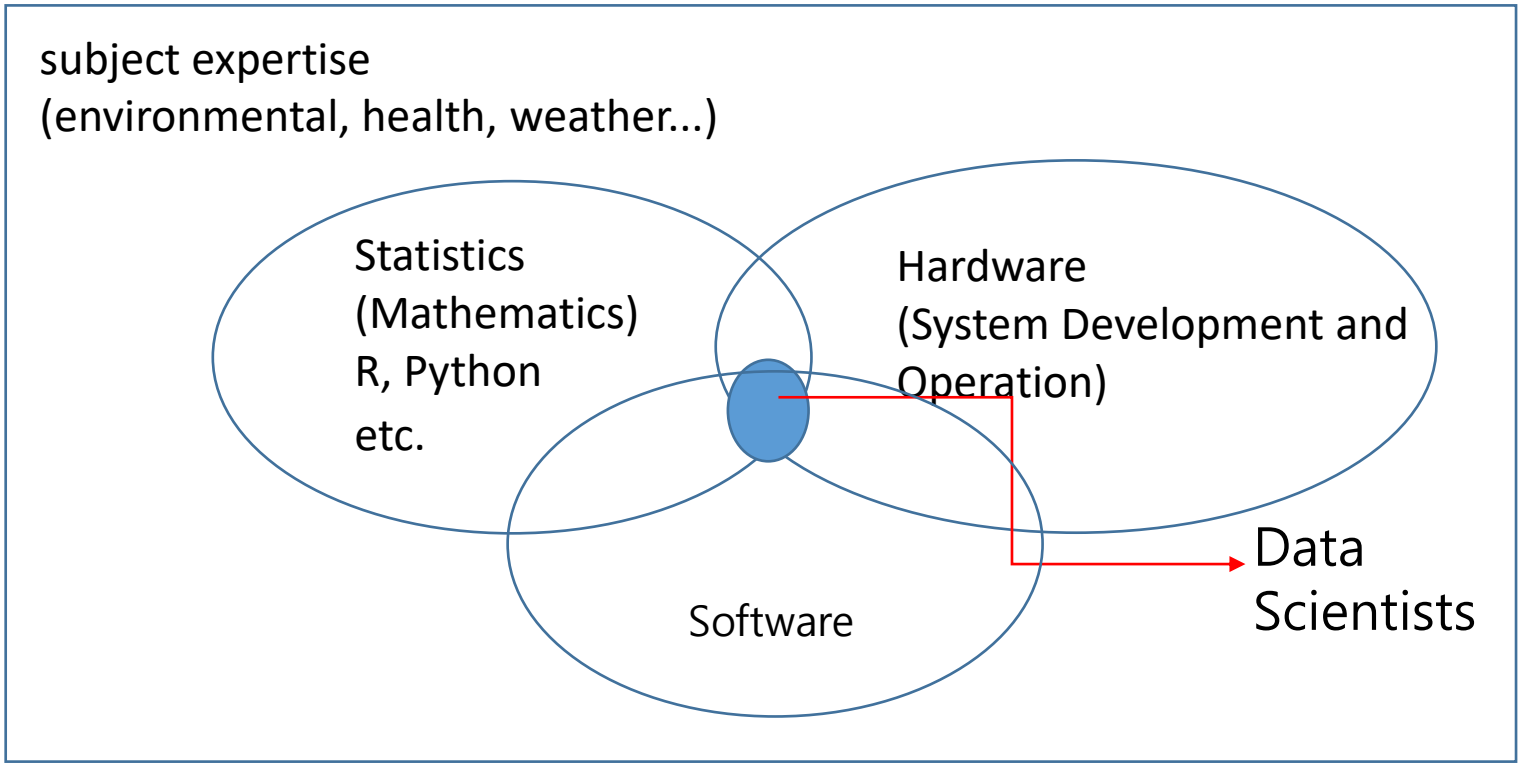
CHAPTER

# Big Data Training Course of STI

1. Background
2. Big Data Program

# Background

How Korea trains statistics-related workers and government officials in the era of big data?



# Background

## Big Data Curriculum Reference Model 1.0

| Essential Coverage | Foundation Coverage | | Technology Coverage | | Analysis Coverage | | Business Coverage | |
|---|---|---|---|---|---|---|---|---|
| Expert | F1 | Insight and Communication | | | | | B1 | Big Data Leadership |
| | | | T1 | Architecture | A1 | Predictive Analysis | B2 | Decision Making and Management |
| | | | | | | | B3 | Policy |
| Advanced | F2 | Persuasion and Negotiation | T2 | Building and utilizing Platforms | A2 | Data Mining | B4 | Project Management |
| | F3 | Logical self-expression | T3 | Processing and Analysis | A3 | Unstructured Data Mining | B5 | Analysis Model and Evaluation |
| | | | | | A4 | Business Intelligence | B6 | Business Analysis |
| Intermediate | F4 | Ethics | T4 | Saving and Management | A5 | Analytical Mind | B7 | Strategic Way of Thinking |
| | F5 | Creative Problem Solving | T5 | Collection | A6 | Statistical Packages | B8 | Big Data Optimization Modeling |
| | | | T6 | Platform Theory | | | B9 | Work Process Knowledge |
| Basic | | | | | A7 | Mathematical Quantitative Thinking | | |
| | F6 | Trend | T7 | Programming | | | B10 | Industry-specific knowledge |
| | F7 | Understanding Business | T8 | Basic IT Theory | A8 | Basic Statistics Theory | B11 | Basic Management Economy knowledge |

Source: Understanding Big Data and Administrative Data (2019), Lecture Material for Statistical Training Institute

| Category | Essential Competency | Explanation |
|---|---|---|
| **Foundation Competency** | Big Data Understanding Business | Ability to understand the correlation between big data and business |
| | Big Data Trends | Ability to understand the latest trends in big data and future developments |
| | Big Data Ethics Consciousness | Realize big data security and privacy |
| | Creative problem solving | Ability to creatively derive big data utilization measures |
| | Insight and Communication | Gain insight into things or phenomena and demonstrate how to use big data |
| | Logical self-expression | Ability to logically express the results of big data utilization |
| | Persuasion and Negotiation | Ability to share and realize big data business opportunities |
| **Platform Technique Competency** | Fundamental IT theory | Ability to understand basic theories and knowledge of IT |
| | Big Data Programming | Ability to leverage the program language used in big data |
| | Big Data Platform theory | Understanding the concepts and key capabilities of the Big Data platform |
| | Big Data Collection | Understand and use the types, methods, and key technologies of big data collection |
| | Big data storage and management | Understand and use the types, methods, and core technologies of big data storage |
| | Big data processing and analysis | Understand and use the types of processing, methods, and key technologies for big data analytics |
| | Big Data Platform Deployment and Utilization | Ability to use big data platform application systems and develop new ones |
| | Big Data Architecture | Ability to design big data IT environments and oversee operations |

| | | |
|---|---|---|
| Analysis Technique | mathematical/quantitative thinking | Ability to interpret problems and derive results based on numbers (quantitative) |
| | analytical mind | Ability to derive new Insight from a wide variety of numbers and statistics |
| | basic statistical theory | Ability to understand basic probabilities/statistics, analytical methods, etc |
| | Utilize statistical packages | Ability to use and apply different statistical packages for different purposes |
| | Business Intelligence | Ability to plan and calculate the analysis results that support management |
| | data mining | Ability to process/analyze data based on knowledge of data mining |
| | Unstructured Data Mining | Understand unstructured data and process/analyze data by each data's characteristic |
| | Big Data Forecast Analysis | Predicting with big data analytics |

| | | |
|---|---|---|
| Business Analytics | basic management/economy knowledge | Ability to understand basic management/economic theoretical knowledge |
| | industry-specific knowledge | Ability to understand industry (manufacturing/distribution/communication, etc.) core business and industry big data strategies |
| | Work process knowledge | Ability to understand/establish business processes (sales/marketing/production, etc.) and big data application strategies |
| | Big Data Optimization Modeling | Ability to establish big data business strategies by industry/business |
| | Business Analytics | Ability to establish analysis direction and application plan in accordance with big data business strategy |
| | Analytical Model and Performance Evaluation | Ability to perform analysis based on analysis strategy and planning results and evaluate the results |
| | Project Management | Ability to manage big data projects to generate targeted performance |
| | Big Data Leadership | Ability to drive business as stakeholders management and general manager of big data business |
| | Decision making and Performance management | Ability to effectively carry out a variety of decisions made during the Big Data business |
| | strategic way of thinking | Understand/improve the correlation between corporate vision/strategies and big data business strategies |
| | big data policy | Understanding ICT and big data policies at home and abroad and applying them to big data businesses |

# Institute

- National Human Resources Development Institute
- Local Government Officials Development Institute
- Korea Human Resources Development Institute for Health and Welfare

General Topics

- Statistics Training Institute → Big Data Analysis

# Big Data vs Traditional Statistics Production

Technology Comparison of Big Data and Traditional Statistical Production Methods

| | Big Data | Traditional Statistical Production |
|---|---|---|
| Data Collection | - Data crawling<br>- Aggregation of various sensors (mobile, traffic, CCTV, etc.) | - Collecting data by hand<br>(Use survey technique)<br>- Manually enter data |
| Data Storage and management | - Distributed storage<br>- Leverage parallel server structure<br>- Automatic data management through the information system<br>- Utilize non-relational databases | - Utilize relational databases<br>- Leverage a single server<br>- Paper material. |
| Data Processing and analysis | - Utilize multiple computer resources through distributed processing<br>- Utilizing Grid computer technology<br>- Active use of advanced analysis technology<br>- Applied to various ranges such as prediction/optimization/ machine learning through data<br>- Utilize packages suitable for processing large amounts of data | - Utilize single computer resources<br>- Using traditional statistical analysis methods |

Source: Understanding Big Data and Administrative Data (2019), Lecture Material for Statistical Training Institute

# Survey Based Training Program vs. Big Data based Training Program

**Survey based Training Program**

- ✓ Introduction to Statistics
- ✓ Advanced Statistics
- ✓ Sample Design
- ✓ Imputation Methods
- ✓ Data Analysis with SAS, SPSS, R
- ✓ Quality Management
- ✓ Statistics Law, Classification

**Big Data based Training Program**

- ✓ Data Analysis with R, SAS, Python, Hadoop…
- ✓ Project-based Data Analysis
- ✓ Deep Learning, Machine Learning
- ✓ Case Study: Mobile data, Online price, Credit cards

# Big Data Training Program

## Understanding Big Data and Administrative Data('17~)

Administrative Data Visualization
Big Data Utilization, Problems and How to Improve
Case Study: Job Statistics Using Administrative Data

## Statistics Production using Administrative data('17)

Current Status of Utilization of Administrative Data and Statistical Utilization Techniques
Preparation of statistics on the utilization of administrative data
Case Study : Hands-on practice of Statistics Production using Newlywed Couple's Administrative Data

## Data- based Policy Making('18~)

Meaning and Importance of Statistical-Based Policies
Statistical issues arising in policy making
Data Analysis Skill for Statistical-Based Administration

## Data Analysis with Python (‘20~)

Basics of Python Programming
Python Data Handling Application, module utilization, data entry and type

- Data Handling : Operator/Date Statement, Operator Utilization, External Data Recalling, Combining
- Basic Statistics : Frequency Analysis and Technical Statistics, Comparison of Two Groups, ANOVA, Chi-Square Test, Correlation Analysis
- Case Study : Data Handling and Statistical Analysis Application Using Marine Accident Data provided by Public Data Portal

## Hadoop-based Big Data Statistical Analysis (‘20~)

Big data platform based on Hadoop
Extracts, preprocesses data from deployed big data platforms
Collecting, storing, processing, and analysis of structured and unstructured data
Utilizing Hadoop Ecosystem, collecting and analyzing real-time data
Statistics Korea Data - Micro Data 2015 Population Census, Small Business Data
Machine Learning - K-NN, Logistic Reform, Decision Tree, Small Business Closure Forecasting Model Development

CHAPTER

# Future Plans

# Future Plans

## Big Data Analysis Specialists on long-term project-type courses

- training on how to shape data by applying statistical analysis in the process of big data collection and pretreatment
- hands-on practice using real data, training experts who can actually apply it to the field
- Case Study

  - Handbook on the use of Mobile Phone Data for Official Statistics

    (UN Global Working Group on Big Data for Official Statistics Draft September(2019))
- Challenges : Data, programs and guidelines on how mobile data is distribute from base station

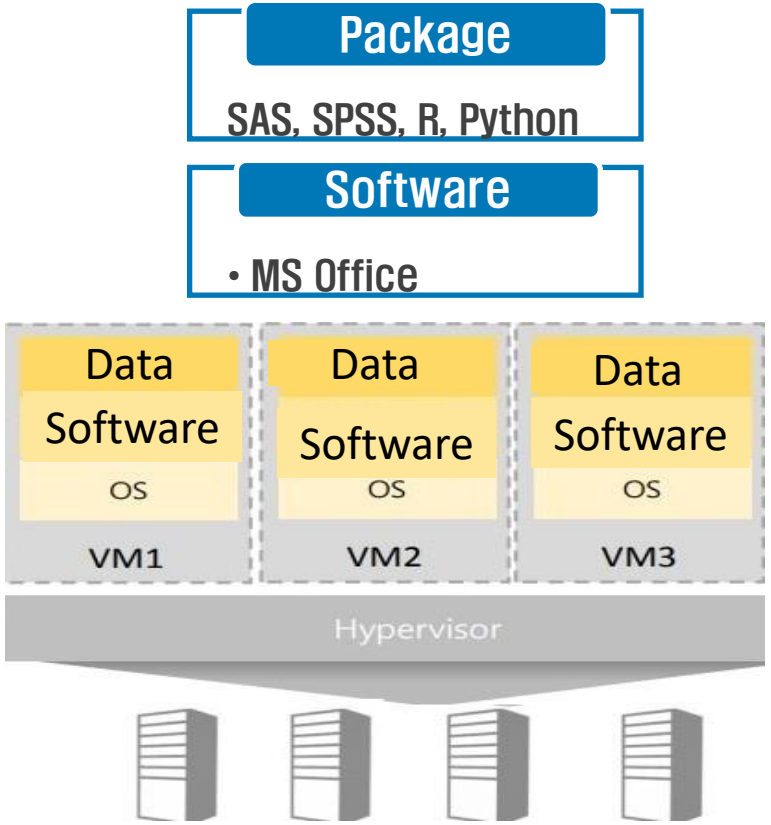
# VDI Desktop

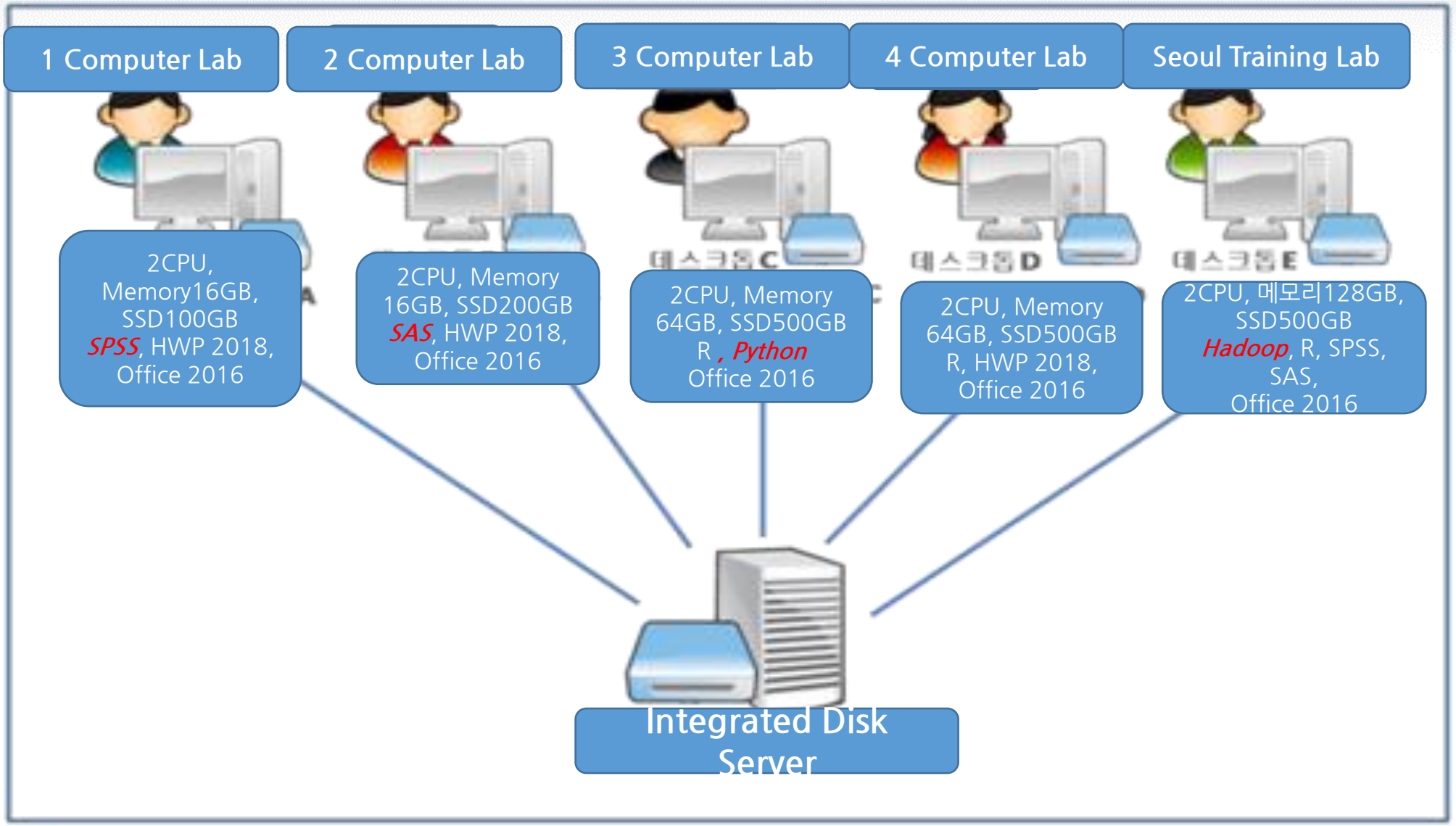
◆ VDI(Virtual Desktop Infrastructure)

- provides customized virtual desktop and storage space for each users using resources on a server

Desktop for training

Zero-Client Computer Terminals
(Unable to save data)




1 Computer Lab
2 Computer Lab
3 Computer Lab
4 Computer Lab
Seoul Training Lab
2CPU, Memory16GB, SSD100GB SPSS, HWP 2018, Office 2016
2CPU, Memory 16GB, SSD200GB SAS, HWP 2018, Office 2016
2CPU, Memory 64GB, SSD500GB R , Python Office 2016
2CPU, Memory 64GB, SSD500GB R, HWP 2018, Office 2016
2CPU, 메모리128GB, SSD500GB Hadoop, R, SPSS, SAS, Office 2016
Integrated Disk Server

# Q & A