



Big Data Platform for Internet Basic Resources and DNS Measurement

Jiankang Yao

CHINA INTERNET NETWORK INFORMATION CENTER

CONTENTS

- 1 Overview
- 2 Data of the Big Data Platform
- 3 Technical capabilities of the Big Data Platform
- 4 Practical Applications

Introduction



By collecting and aggregating data of Internet basic resources in an all-round way, the Big Data Platform of National Internet Basic Resources aims at supporting the comprehensive verification and analysis of the security and operation of network basic resources and displaying the operation and development trends of China's Internet. It provides solid data support, platform support and support for the capacity of analysis and mining in many fields, including the research on the status, the future development and fundamental security of the Internet.

Three Phases

- The **first** phase of the platform project focuses on the aggregation, sorting, and analysis of CNNIC own data resources, as well as initially establishing an external data collection system, providing strong support for measurement and analysis of domestic key domain names and DNS resolution systems.
- The **second** phase of the platform project focuses on the support of big data platforms for key business work of relevant ministries.
- The **third** phase of the platform project aims to achieve sustainable development of the platform, at the same time provide public and customized services in the field of internet basic resources.

Construction progress

- Phase I, released on September 20, 2018.
- Phase II, released on November 15, 2020.
- Phase III was launched in 2021. It is being carried out in an orderly manner as planned.

Unique Data Assets

Passive data

Sort out, connect, aggregate, integrate, clean, and warehouse data from CNNIC business systems

- More than **20 million** “.CN” domain name registration data.
- Continuously growing domain name resolution data, including over **10.2 billion** national top-level authoritative domain name resolution data per day, and 1.2.4.8 public recursive DNS service resolution data for public services.
- Accounting for **38.78%** (24483/32 allocated), **25%** (86005760 allocated), and **17.8%** (1113 allocated) of **IPv6**, **IPv4**, and **AS** number allocation data in mainland China, respectively
- **Statistical and investigative data** on the development of China's Internet, including multiple dimensions such as the construction of internet basic resources, the size and structure of internet users, and the development of internet applications.



Global Data
Collection



Active Data
Collection



Big Data Platform

Global integrated data
collection and
measurement

Data Processing and
Analysis Based on
Graph Neural Network

Knowledge Graph
of Internet Basic
Resources

Data security
guarantee based on
privacy computing

- More than 100 DNS measurement nodes.
- Covering over 1100 TLDs with over 230 million domain names.
- Routine measurement and analysis of global domain name applications and harmful websites, domain name resolution, IPv6 applications, etc.
- Establish an internet basic resource data indicator system covering 38 major categories and 74 subcategories.
- The total size of data assets amounts to 420TB.

Data Asset Scale

- Build a global internet infrastructure resource database, with accumulative data volume of 2539.2 billion pieces.

Supporting the management of China's basic internet resources

- Continuously Gathering Internet Basic Resource Data
- Supporting the Ministry to grasp internet basic data and management
- Establish management support for our department's compliance service capabilities, technical operation capabilities, and operational management capabilities for internet basic resource practitioners
- Strengthen data integration to provide support for multi-dimensional

Supporting dynamic mastery of global internet basic resources

- A global internet infrastructure surveying and mapping system, timely perceive and discover the operation and security situation.

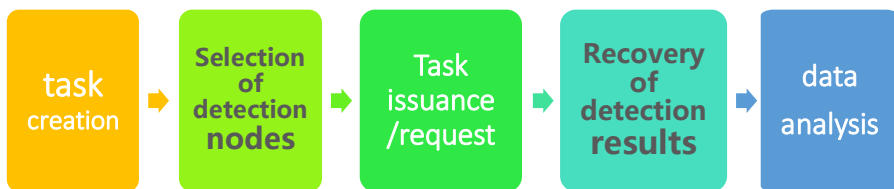
Domain name resolution service measurement and evaluation

- Routine measurement and statistics;
- Conduct a comprehensive evaluation (authoritative and recursive) around security, performance, traffic, faults, configuration, etc.;
- Deep measurement of TLD operations in global relevant countries during hot events.



The measurement and analysis of root domain, TLD, and second level and lower domain names under China's domain name service system, as well as recursive domain name services, cover automated measurement of key indicators related to domain name service and security, such as NS configuration, service latency trend, port randomness, TCP/EDNS0/IPv6 support, DNSSEC, BIND version, DOT support, DOH

类别	监测项	监测指标	说明	是否支持周期探测	探测度
性能/故障	可用性	NS记录	域名权威服务器记录	是	可配
		NS IPv4	域名权威服务器记录对应的IPv4地址	是	可配
		响应时间	域名服务器DNS查询响应时间	是	可配
		SOA序列号	区SOA记录中的序列号值	是	可配
		SOA记录	否	否	否
	A记录探测	可用性	域名服务器DNS查询服务是否可用	是	可配
		A记录	域名对应的A记录	是	可配
		从指定递归获得的A记录TTL	指定递归回复的A记录的TTL值	是	可配
		HTTP服务响应码	域名A记录对应的HTTP服务响应码	是	可配
		解析时间	A记录查询响应时间	是	可配
安全	DNSSEC支持性	是否支持DNSSEC	是否支持DNS安全协议 (DNSSEC)	是	可配
	NSEC3支持性	是否支持NSEC3	是否支持NSEC3协议 (NSEC3是为那些不存在的域名设定的, 它能在部分程度上抵御区记录被枚举)	是	可配
		DNSKEY记录	域名DNSKEY记录 (DNSSEC中的公钥)	是	可配
	DNSKEY	DS记录	域名DS记录 (DNSSEC公钥的签名)	是	可配
		DS验证是否通过	DS信任链验证是否通过 (签名是否验证通过)	否	否
	递归查询支持性	是否支持递归查询	权威服务器是否支持递归查询	是	可配
	端口随机性探测	递归服务器进行迭代查询时所使用的源端口的随机性程度	是	是	可配
		递归服务器迭代查询源端口随机性标准差值	是	是	可配
	区域传送支持性	区域传送支持性	域名是否支持区数据传送	是	可配
	链路丢包率探测	链路丢包率探测	探测点至域名服务器的链路丢包率 (借助ping进行)	是	可配
流量	traceroute	traceroute	探测点至域名服务器的路由信息	否	否



- **Data Tasks** : Periodic、real-time、mass tasks
- **Detection Priority** : real-time > Periodic > mass
Prioritize the execution of high priority tasks
- **Fault Tolerance** : Heartbeat mechanism detects detection nodes, dynamically allocates detection tasks, and locally caches and uploads result data in blocks.

类别	监测项	监测指标	说明	是否支持周期探测	探测度
配置	主辅一致性	主辅一致性	域名各权威服务器是否一致 (SOA序列号是否一致)	是	可配
	NS配置正确性	NS配置正确性	域名NS记录是否配置正确 (各权威服务器NS记录是否一致)	是	可配
	NS冗余性	冗余性	权威服务器冗余性 (NS数量)	是	可配
	BIND版本应答支持性	BIND版本应答支持性	是否支持BIND版本应答特性 (通过发ch txt version.bind查询包判定)	是	可配
	BIND应答字符串	BIND版本应答回复字符串	是	是	可配
	TTL	SOA TTL	域名SOA记录TTL值	是	可配
	通配符支持性	通配符支持性	域名服务器是否支持通配符查询 (通过查询不存在的域名进行判定, 不存在的域名借助guid构造)	是	可配
	TCP查询支持性	是否支持TCP查询	域名服务是否支持以TCP方式进行查询	是	可配
	IPv6支持性	NS IPv6	域名权威服务器对应的IPv6地址	是	可配
		NS IPv6是否可达	域名权威服务器IPv6地址是否可达	是	可配
	EDNS0	是否支持EDNS0	是否支持DNS扩展版本0协议; EDNS0是为了解决原有DNS协议的一些限制, 如只能发送不超过512个字节DNS包的限制	是	可配
		EDNS0最大包长	服务器端声称的支持的EDNS0最大包长值	是	可配
	DNS解析软件探测	DNS解析软件探测	域名服务器解析软件类型和版本探测 (借助fpdns进行)	是	可配
	操作系统探测	操作系统探测	域名服务器操作系统类型和版本探测 (借助nmap进行)	是	可配
		开放端口探测	域名服务器开放端口探测 (借助nmap进行)	是	可配
网站探测	是否存在网站应用	是否存在网站应用	域名对应A记录服务器是否部署有网页服务	否	否
	时钟偏差探测	时钟偏差探测	探测点至域名服务器的时钟偏差情况 (基于TSIG协议进行)	否	否
	授权一致性探测	上下级授权是否一致	域名本身的NS记录和上级授权的NS记录是否一致	是	可配
		NS TTL	域名NS记录的TTL值	否	否
		父域获得的本地NS TTL	授权NS记录的TTL值	否	否
	MX	MX记录	域名的邮件记录	否	否
		MX IPv4	域名邮件记录对应的IPv4地址	否	否
		MX IPv4 PTR	域名邮件记录IPv4地址对应的反向记录	否	否
	镜像ID探测	镜像ID响应时间	Anycast镜像ID查询的响应时间	是	可配
		镜像ID	Anycast镜像ID信息 (通过发ch txt hostname.bind包获取)	是	可配
IPV6	最大回复包长检测	最大回复包长检测 (实测模式)	实际测得的递归服务器所支持的EDNS0最大包长	否	否
	递归后端IP	递归后端IP探测	递归进行迭代查询时所使用的后端IP	否	否
	ECS支持性	ECS支持性	ECS协议的支持性, ECS协议在查询过程中会发送用户IP, 以加速用户端访问	否	否
	Web IPv6支持性		查询域名IPv6地址, 并进行Web访问支持性探测	是	可配
MX	MX IPv6支持性		查询域名MX记录及MX IPv6地址记录, 并进行邮件传输协议 (SMTP) IPv6支持性探测	是	可配
	DNS IPv6支持性		查询域名NS记录及NS IPv6地址, 并进行NS解析IPv6支持性探测	是	可配
网站探测		中国全部网站数	中国网站数量探测, 包括ip地址在国内的网站和所有国家域名对应网站。	是	否

3

Technical capabilities of the Big Data Platform

Full chain technology system in the field of big data

- Key technology R&D , focused on the field of internet basic resources
- layer by layer processing and block management; research and design modules for data collection, cleaning, processing, mining, openness, application, and management; forming a customizable and scalable technology architecture.

Qualified personnel team of big data technology

- Established a technical team and trained a group of technical talents
- A R&D team familiar with the characteristics of internet basic resource data & cloud computing & deep learning

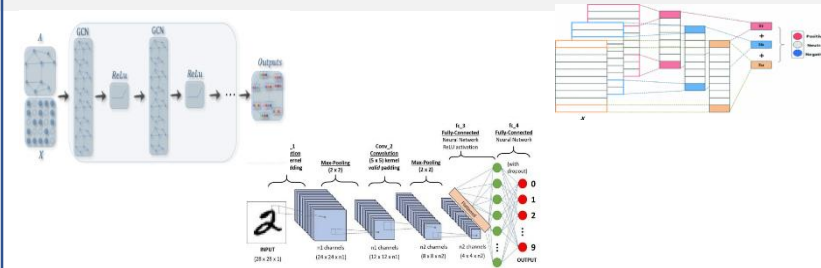
Intellectual Property & Related Standards

- 5 national industry standards, published more than 10 academic papers, and obtained more than 20 national invention patent authorizations.



Large scale data analysis and model training capabilities

- Capable of large-scale data analysis and deep learning model training
- Completed large-scale data-intensive analysis and high-performance GPU computing analysis environment
- Key technologies, such as domain name graph embedding representation analysis for analytical relationships and domain name multiple graph embedding representation analysis based on self attention mechanism
- Established the ability to service big data applications such as domain name detection and detection of malicious applications and security situation analysis through the knowledge graph of internet basic resources.



Ministry of Science and Technology - New Domain Name Resolution System

- Design and Development of a Novel Domain Name Resolution Prototype System using Blockchain
- Real time measurement of goals such as co governance root, co governance chain, enhanced recursion, and experimental TLD.
- Stable operation since the system was launched.



Ministry of Science and Technology - Key Information Analysis Technology for Internet Infrastructure

- The overall goal of the project is to carry out perception, analysis, service, and demonstration verification of key information in internet infrastructure.
- CNNIC undertakes Task III of the project, mainly researching key information analysis technologies for internet infrastructure, and achieving the fusion analysis of internet infrastructure data.
- 5 sub tasks:
 - Research on DNSSEC and SSL/TLS certificate situational awareness technology ;
 - Research on the Construction Technology of Knowledge Graph for Internet Infrastructure;
 - Research on Perception and Analysis of Internet Infrastructure Security Situation;
 - Research on the Index System for Internet Infrastructure Security Risk Assessment;
 - Research on Internet Infrastructure Data Security Management System.
- Execution period: Dec. 2022 to Nov. 2025

4

Application Effectiveness

**A national grade
internet basic
resources
management
platform**

measurement and
Disposal of malicious
domain names

IPv6 Application
deployment measurement

Key protection domain
name security
measurement

**Internet Basic Resource
Situation Awareness**

Statistical Report on the
Development of China's
Internet

Annual Analysis Report on the
Security of China's Domain
Names

Monthly report

Data Analysis of China's
Domain Names

**Special Reports for ministries
and commissions(data analysis)**

**Global Internet Basic Resources
Mapping**

System Support

Data Support

Data mining analysis

**Global Internet Basic
Resources Database**

Big Data (Service) Platform of National Internet Basic Resources

**Active
probing
data**

**DNS Detection
data**

**Website
Detection data**

**data
aggregation**

**Passive
data**

**DNS Authoritative
logs**

**DNS Recursive
logs**

**CN registration
data**

Practical Applications

—— Internet Basic Resource Security Situation Awareness

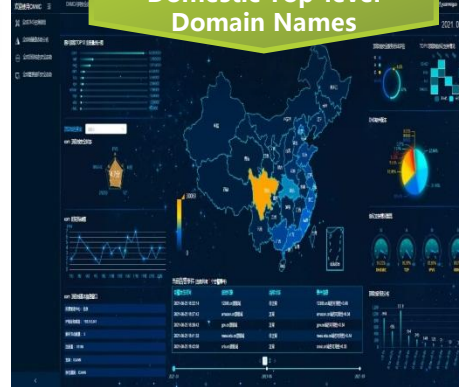
Domestic Root Mirror
Security Situation



Security Situation of
Domestic Authoritative
Domain Names



Security Situation of
Domestic Top-level
Domain Names



Global DNS
Asset Mapping



CN National Domain
Name Security Situation



Key Domain Name
Security measurement



Internet Basic Resource Security Situation Awareness

- Gathering, integrating, and detecting important internet basic resources data such as domain names, DNS resolution, and BGP, etc.
- Sensing and discovering the basic network security status, identifying network security events and operational failures of key measurement notes.
- Mastering the security situation and threat intelligence clues of IBR, providing early warning, outputting relational data measurement and analysis reports.

Domain Name System Data Assurance Service

- Provide Real time operational status measurement & warning and big data dashboard for national domain names 24-hour safe and stable operation.
- Safety measurement, Early Warning report, Statistical Analysis, Guarantee support.
- Carry out statistical analysis : overseas registration& application and website

Network security guarantee for major events

- Global distributed measurement of the service status, security situation, phishing and phishing of key protected domain names
- Important guarantees tasks: National Day, World Internet Conference, Beijing Winter Olympics, Winter Paralympics, Epidemic Prevention and Control, 20th National Congress and many other events.
- Support Annual Analysis Report on the Security of China's Domain Names.

Serve Data Basis

- 13 global root and root mirror servers, over 1 million DNS servers, and real-time detection and perception of the working status of public DNS servers.
- Data: IP address, geographic information, DNS version, security configuration, response time, and service status and so on.

Analyze the concentration of key domain name resolution servers, Determine whether they are concentrated on specific domain name servers to Avoid query pressure caused by measurement .

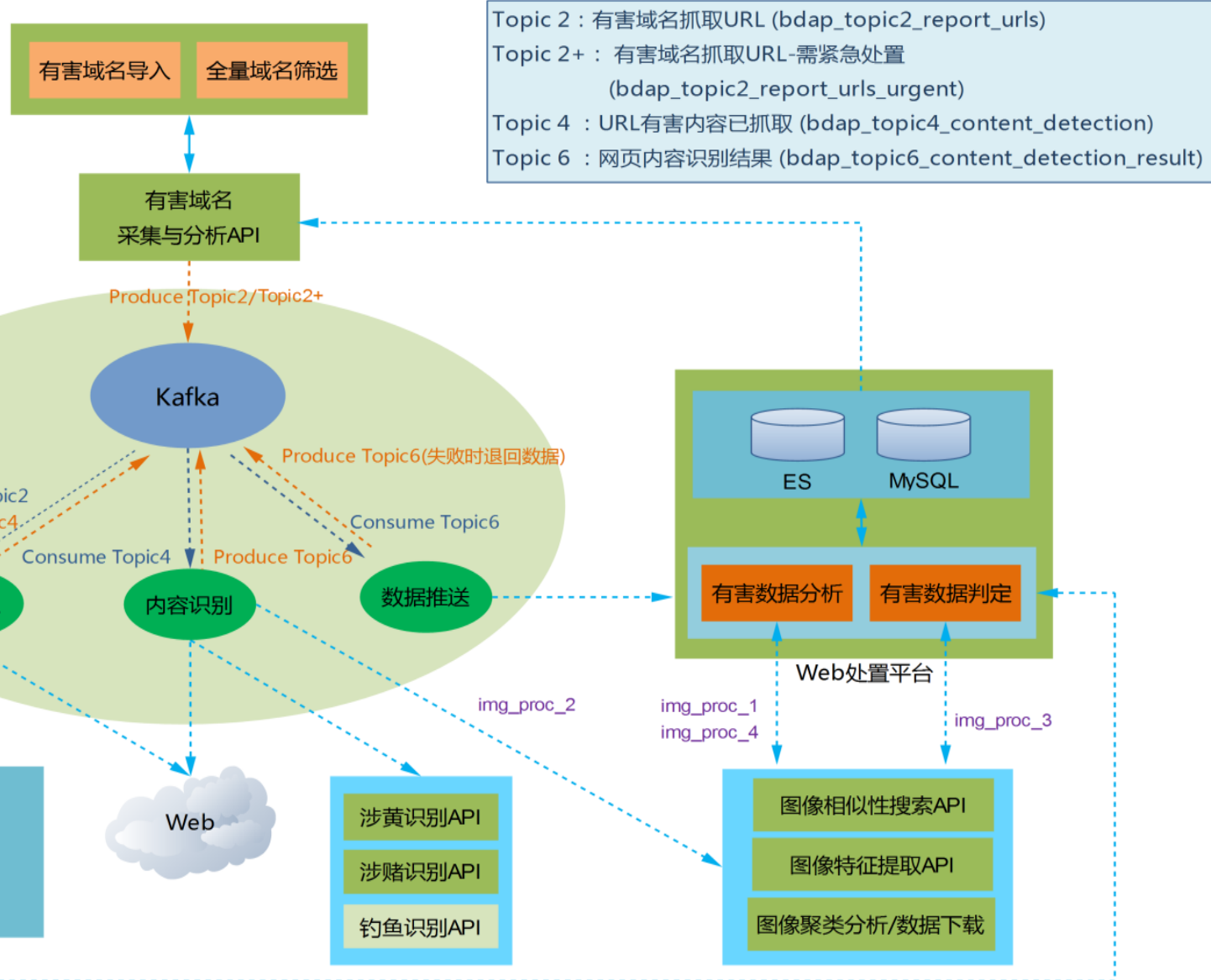
Focusing on the geographical location of key domain name servers, four preferred geographical directions of east, south, west, and north are selected for detection nodes.

For key domain names with highly centralized resolution, increase the detection interval appropriately to avoid query pressure on the domain server.

Practical Applications

— measurement and Disposal of malicious domain names

img_proc_1 : 图像特征库下载
img_proc_2 : 图像相似性搜索
img_proc_3 : 待聚类图像集下载
 图像聚类结果提交
img_proc_4 : 图像特征提取



- Since 2023, measurement has found **3.0056 million** suspected cases of involvement in Suspected **Phishing** related which has exceeded 90 points by machine learning Scoring. In which, approximately **285500 names** in the ccTLD.
- In 2023, Rely on sub platform supporting the **disposal** of **300 thousand** malicious applications of national domain names.

Practical Applications

— measurement and Disposal of malicious domain names

- Within one week, it is possible to complete in-depth detection and identification analysis of phishing websites in ".CN" and over 1100 universal top-level domains totaling 230 million domain names.

➤ Multiple round Screening Strategy

➤ Global Domain Name Scanning

Large scale universe **detection** and **Quick screening** based on domain name- IP-AS mapping

- Quickly find a large number of potential malicious domain names by using a small number of known malicious seeds.
- Experiments show ,quick screening of 15000 level domain names "domain name- IP- AS" features can be achieved by using around 200 harmful domain name seeds.

Harmful text recognition based on Machine learning

- A Suspected and Fast malicious Domain Name Detection Algorithm based on Text Training Model Recognition and Graph Analysis
- A single server can quickly filter harmful content from text-based websites for 20 million domain names in a single day.

Harmful Image recognition based on machine learning

- Use screenshot images of suspected harmful application domains and phishing websites as search items.
- Can complete image deep recognition detection based on deep learning for approximately 200000 harmful domain name websites based on images in a single day.
- Train and compare the identified harmful screenshots to improve the efficiency of harmful domain name review.

Overview of Global malicious Application Domain Name Data



Overview of China's malicious Application Domain Name Data



Distribution of malicious Application Domain Names



Practical Applications

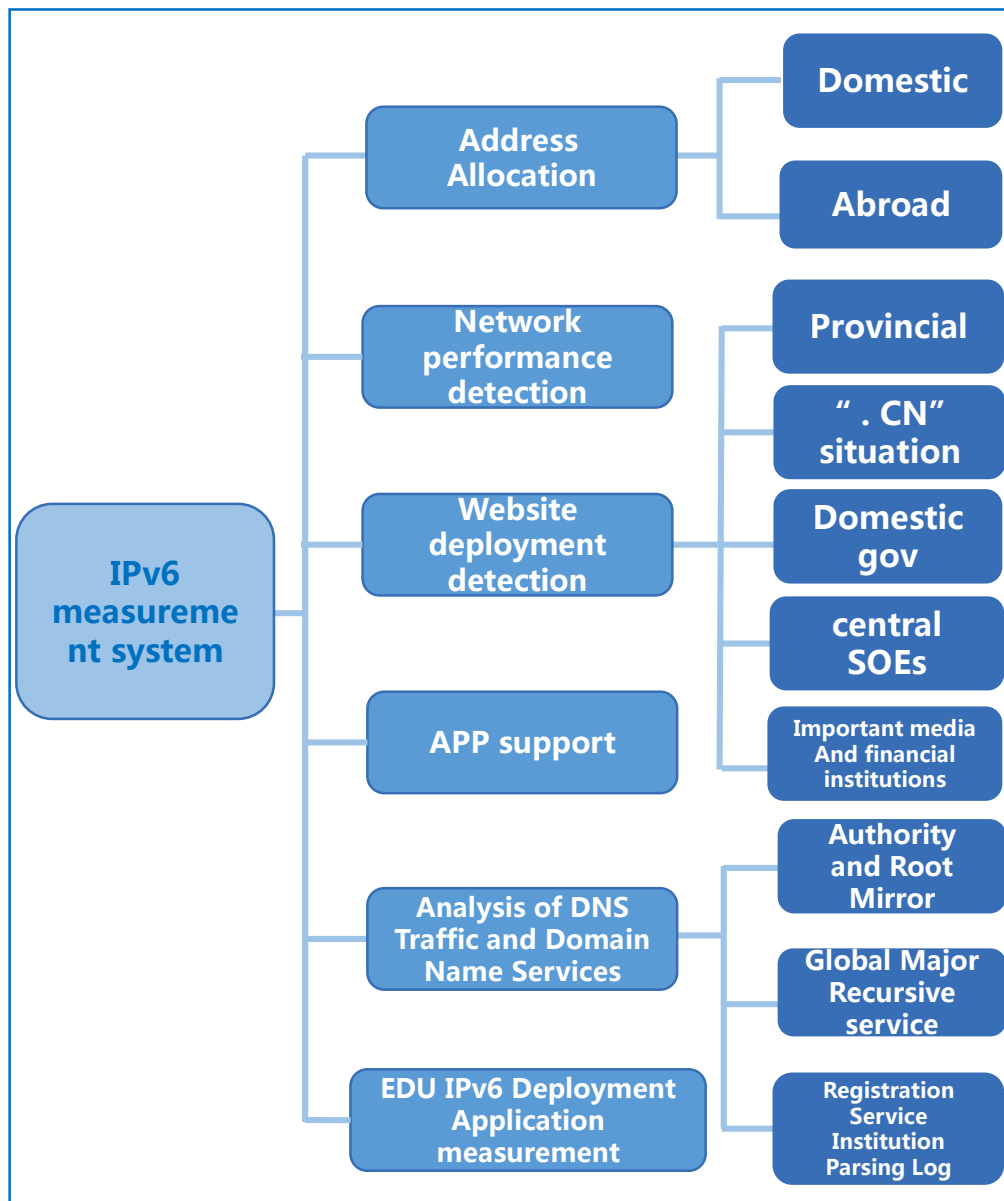
— IPv6 Development measurement

National Policy for Promoting IPv6

- 《 Action Plan for Promoting the Scale Deployment of Internet Protocol Version 6 (IPv6) 》 ,
- 《 Three Year Special Action Plan for IPv6 Traffic Enhancement (2021-2023) 》

- Relying on the integrated collection and analysis of big data platforms and the advantages of CNNIC's internet basic resource data, we construct an IPv6 measurement index system that comprehensively reflects the deployment of IPv6 applications in China.

- Continuously improve the development of IPv6 application measurement



Practical Applications —— Supporting Internet Basic Resource Management

Timely perceive and discover the latest security situation of global internet basic resources

- timely perceive and discover the latest security situation of global internet basic resources, and ensure the basic security of China's digital economy development
 - DNS-OARC Community related shared data
 - Regular operation and security measurement of global roots (including mirrors), 1100 TLDs worldwide, over 1 million DNS resolution servers worldwide, and key websites in important countries & regions.

Conduct in-depth analysis and research on the development trends of the global internet & digital economy

- Relying on the advantages of platform data resources and key technologies for large-scale data analysis, based on relevant analysis of the development of global internet basic resources, we will conduct forward-looking research and judgment on the development trends of the global internet and digital economy, and plan and respond in advance.





Thanks!